A corpus-driven study of features of Chinese students' undergraduate writing in UK universities

Thesis

A CORPUS-DRIVEN STUDY OF FEATURES OF CHINESE STUDENTS' UNDERGRADUATE WRITING IN UK UNIVERSITIES

Maria Elizabeth Leedham

(BA, MSc)

Thesis submitted to the Open University in fulfillment of the requirements for the degree of Doctor of Philosophy
Faculty of Education and Language Studies

MAY 2011

ABSTRACT

Chinese people now comprise the 'largest single overseas student group in the UK' with more than 85,000 Chinese students registered at UK institutions in 2009 (British Council, 2010a). While there have been many studies carried out on short argumentative essays from this group (e.g. Chen, 2009), and on postgraduate theses (e.g. Hyland, 2008b), there has been comparatively little research conducted on the high-stakes genre of undergraduate assignments. This study examines assessed writing from Chinese and British undergraduates studying in UK universities between 2000 and 2008; these are investigated using corpus linguistic procedures, supported by qualitative reading.

A particular focus is the use of lexical chunks, or recurring strings of words. Findings from the literature on Chinese students' written English indicate high use of informal chunks, connecting chunks, and those containing first person pronouns (e.g. Milton, 1999). This study found that while the Chinese students make greater use of particular connectors and the first person plural, both student groups make (limited) use of informal language. These areas of difference are more apparent in year 1/2 assignments than those from year 3, suggesting that students gradually conform to the academy's expectations. Unexpected findings which have not been previously identified in the literature include Chinese students' significantly higher use of tables, figures (or 'visuals') and lists, compared to the British students' writing. Detailed exploration of writing within Biology, Economics and Engineering suggests that using visuals and lists are different, yet equally acceptable, ways of writing assignments.

Since the writing of both student groups has been judged by discipline specialists to be of a high standard, it is argued that the difference in use of visuals and lists illustrates the range of acceptability at undergraduate level. The thesis proposes that scholars therefore need to consider expanding the notion of what constitutes 'good' student writing.

ACKNOWLEDGEMENTS

This thesis owes a great deal to my supervisors, Ann Hewings, Barbara Mayor, and Sarah North: thank you for your patience and constant encouragement over the last four years. I'm also grateful to The Open University for granting me a studentship for full time PhD study, and to my external examiners, Paul Thompson and Alison Wray, for their support and contributions. The study would not have been possible without the participation of the Chinese and British students in the BAWE project, and I'd like to thank the friends and colleagues who encouraged their students to submit additional assignments and to respond to my questionnaire. I'm grateful to the people who read and commented on chapters from the thesis (Lina Adinolfi, Guozhi Cai, Lynne Cameron, Signe Ebeling, Kieran O'Halloran, Prithvi Shrestha and Pete Whitelock), and to my Dad and Aqsa Dar for the proofreading. Finally, thank you to my family for their continuing support: my husband, Pete Whitelock, for all the cooking; my sons, Bob, Jim and Matt, for putting up with a distracted mother; and to my parents, John and Viv Leedham, for their confidence in me.

Note: The data in this study come from the British Academic Written English (BAWE) corpus, which was developed at the Universities of Warwick, Reading and Oxford Brookes under the directorship of Hilary Nesi and Sheena Gardner (formerly of the Centre for Applied Linguistics [previously called CELTE], Warwick), Paul Thompson (formerly of the Department of Applied Linguistics, Reading) and Paul Wickens (Westminster Institute of Education, Oxford Brookes), with funding from the ESRC (RES-000-23-0800).

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION 15

1.1 Goal of the thesis 15

1.2 Chinese students 17

1.2.1 Numbers of Chinese students in the UK 17
1.2.2 Commonalities of Chinese students 19
1.2.3 Educational background of students from the PRC 23

1.3 Challenges in assignment-writing 30

1.3.1 Challenges for all students 30
1.3.2 Additional challenges for Chinese students 32

1.4 Organization of the thesis 34

1.5 Chapter summary 36

CHAPTER 2 CHINESE STUDENTS' WRITING IN THE LITERATURE 39

2.1 Initial review questions 39

2.2 English students' undergraduate writing 42

2.3 Characteristics of Chinese students' writing 45

2.4 Variation over time 56

2.5 Disciplinarity 64

2.6 Discussion and implications 73

2.6.1 Deficit versus descriptive perspectives 74
2.6.2 Comparison of ICLE and BAWE 75

2.7 The research questions 84

2.8 Chapter summary 86

CHAPTER 3 ANALYZING STUDENT WRITING THROUGH A FOCUS ON LEXICAL CHUNKS 89

3.1 Introduction 89

3.2 Relating lexical chunks to student writing 91

3.3 Identifying lexical chunks 102

3.3.1 Characteristics of lexical chunks 103
3.3.2 Methods of identifying chunks 106

3.4 The view of lexical chunks adopted in this study 111

3.5 Classification of chunks 117

3.5.1 Structural classification 118
3.5.2 Functional classification 120

3.6 Chapter summary 125

CHAPTER 4 DATA AND RESEARCH METHODS IN THE STUDY..... 127

4.1 Introduction 127

4.2 The data 129

4.2.1 The BAWE corpus 130
4.2.2 Additional assignments 131
4.2.3 Refining the corpora 134
4.2.4 The final data 138

4.3 Corpus linguistic procedures 144

4.3.1 Describing characteristics of the texts 145
4.3.2 Extracting keywords 147
4.3.3 Counting n-gram tokens and n-gram types 152
4.3.4 Analyzing 4-grams 159

4.4 Text level analysis 161

4.5 Chapter summary 163

CHAPTER 5 FEATURES OF CHINESE STUDENTS' WRITING IN THE CORPUS.....165

5.1 Introduction 165

5.2 Text characteristics 166

5.3 Keyword analysis 170

5.3.1 Informal items 171
5.3.2 Connectors 177

5.3.3 First person pronouns 181
5.3.4 References to data and visuals 191

5.4 Chapter summary 196

CHAPTER 6 VARIATION ACROSS YEAR GROUPS..... 199

6.1 Introduction 199

6.2 Text characteristics 200

6.3 Variation in the identified characteristics across the year groups 203

6.3.1 Informal items 204
6.3.2 Connectors 204
6.3.3 First person pronouns 208
6.3.4 Visuals and lists 211

6.4 Textbook study 212

6.5 N-gram tokens 218

6.6 Classification of four-grams 221

6.6.1 Structural classification 221
6.6.2 Functional classification 224

6.7 Chapter summary 227

CHAPTER 7 DISCIPLINARY INFLUENCES 229

7.1 Introduction 229

7.2 Keyword analysis 230

7.2.1 The data 230
7.2.2 Student writing in Biology, Economics and Engineering 231
7.2.3 Connectors 235
7.2.4 First person pronouns 236
7.2.5 Visuals and lists 237

7.3 Visuals and lists: Whole text analysis 239

7.3.1 Visuals and extended captions in Biology 239
7.3.2 Bulleted lists vs. connected prose in Economics 244
7.3.3 Formulae and white space in Engineering 248
7.3.4 Interviews with lecturers 251

7.4 Conclusion and summary 254

CHAPTER 8 CONCLUSIONS 257

8.1 Answering the research questions 258

8.2 Implications 2632

8.3 Limitations and suggestions for future research 266

REFERENCES 271

APPENDIX A 293

APPENDIX B 295

APPENDIX C 297

APPENDIX D 300

APPENDIX E……303

APPENDIX F……305

APPENDIX G……308

LIST OF FIGURES

Figure 1.1 University English language class....25

Figure 2.1 Soft-hard, pure-applied discipline paradigm 65

Figure 2.2 Sample instructions for English composition paper 81

Figure 3.1 *that there is a* 115

Figure 3.2 Overview of lexical chunks 116

Figure 4.1 Disciplines in the study....142

Figure 4.2 Comparison of normalized token counts of 4-grams 155

Figure 4.3 Comparison of normalized token counts of 4-grams 156

Figure 5.1 *a little bit* in Chi123 173

Figure 5.2 Plot dispersion for *I* in Chi123....188

Figure 5.3 Plot dispersion for *I* in Eng123....189

Figure 5.4 Extract from 0254j showing integration of formulae....193

Figure 5.5 Example of listlike in Engineering 194

Figure 6.1 Informal items across the year groups....204

Figure 6.2 Variation in use of connectors across year groups....205

Figure 6.3 Variation in use of however and therefore across year groups....206

Figure 6.4 Number of visuals and lists by year group....211

Figure 6.5 Example One: List of connectors 213

Figure 6.6 Example Two: Question from Fukian NMET, 2005....215

Figure 6.7 Example Three: Question from Sichuan NMET, 2006....216

Figure 6.8 Comparison of n-gram tokens in the four corpora 219

Figure 6.9 Structural categorization using broad VP, PP, NP groupings 222

Figure 6.10 Broad functional categorization of 4-grams 224

Figure 7.1 Page 1 of Biology assignments 241

Figure 7.2 Diagrams and extended caption in 0434a p.4 (Chinese writer) 242

Figure 7.3 Visual and extended caption in 0434a p.5 (Chinese writer)....243

Figure 7.4 Example page of each Economics assignment 245

Figure 7.5 Extract from text 0155a (Chinese writer) 246

Figure 7.6 Extract from text 0202j 247
Figure 7.7 Discussion/evaluation sections in two Engineering assignments 249
Figure 7.8 ‘Conclusions’ in two Engineering assignments 250

LIST OF TABLES

Table 1.1 All non-UK domiciled students in Higher Education in 2008/9 18

Table 1.2 Top 10 non-EU senders of students to UK HEIs in 2008/9 19

Table 2.1 Comparison of learner corpus texts and authentic undergraduate assignments: conditions of writing 77

Table 2.2 Comparison of learner corpus texts and authentic undergraduate assignments: Factors specific to students 78

Table 3.1 Structural classification of chunks 119

Table 3.2 Participant-oriented metafunctions 122

Table 3.3 Research-oriented metafunctions 122

Table 3.4 Text-oriented metafunctions 123

Table 4.1 Number of tokens and texts per student corpus (before refining) 133

Table 4.2 Progressive refinement of datasets 137

Table 4.3 Number of tokens and texts per student corpus 138

Table 4.4 Number of tokens and texts per year group corpus 138

Table 4.5 Wordcounts per discipline as a percentage of each corpus 140

Table 4.6 Comparison of software calculations of MSL 146

Table 4.7 Summary of frequency and dispersion thresholds 153

Table 4.8 4-grams in Chi123 (20pmw, 10% texts) 156

Table 4.9 Comparative normalization of types 158

Table 5.1 Descriptive statistics for Chi123 and Eng123 166

Table 5.2 Top 20 negative keywords in Chi123 168

Table 5.3 Key words containing informal items in Chi123 171

Table 5.4 Key connectors in Chi123 178

Table 5.5 Key words containing informal items in Chi123 173

Table 5.6 Keywords containing first person plural in Chi123 182

Table 5.7 Pronoun use in the two corpora 183

Table 5.8 Classification of functions of *we* and *I* in 100 random lines 186

Table 5.9 Keywords containing references to data in Chi123 192

Table 5.10 Counts of visuals and lists in the two corpora 195
Table 6.1 Descriptive statistics for Chi12, Chi3, Eng12 and Eng3 200
Table 6.2 Wordlengths in Chi12 and Chi3 202
Table 6.3 First person pronouns across four corpora 208
Table 6.4 Classification of functions of *we* in four corpora 209
Table 6.5 Classification of functions of *I* in four corpora 209
Table 6.6 Structural classification of chunks 223
Table 6.7 Functional categorization of n-grams 226
Table 7.1 Texts and wordcounts in each discipline subcorpus 231
Table 7.2 Keywords in three disciplines 231
Table 7.3 Collocates of *were* in Biology 232
Table 7.4 N-grams containing *we* in Economics 242
Table 7.5 Connectors in Economics (per 10,000 words) 235
Table 7.6 Statistical comparison of first person pronouns used by each 236
Table 7.7 Use of tables, figures, lists and listlikes per 10,000 words 238
Table 7.8 Comparison of two Biology assignments 240
Table 7.9 Comparison of two Economics assignments 244
Table 7.10 Comparison of two Engineering assignments 248

LIST OF KEY ABBREVIATIONS IN THE THESIS

BAWE	British Academic Written English (corpus project)	**L1**	First language
BNC	British National Corpus	**L2**	Second or additional language
CHC	Confucian Heritage Culture	**LGSWE**	Longman Grammar of Spoken and Written English
Chi123	Corpus of first language Chinese students' writing from undergraduate years 1, 2, and 3	**MSL**	Mean Sentence Length
EAP	English for Academic Purposes	**MWL**	Mean Word Length
EFL	English as a Foreign Language	**NNS**	Non Native Speaker
ELT	English Language Teaching	**NS**	Native Speaker
Eng123	Corpus of first language English students' writing from undergraduate years 1, 2, and 3	**pmw**	per million words
FEI	Fixed Expressions and Idioms (Moon, 1998a)	**PRC**	People's Republic of China
HE(I)	Higher Education (Institution) (HEI is used here as a synonym for 'university')	**RC**	Reference Corpus
HLTM	Hospitality, Leisure, and Tourism Management	**RQ**	Research Question
ICLE	International Corpus of Learner English	**TESOL**	Teaching English to Speakers of Other Languages
IELTS	International English Language Testing System	**ToEFL**	Test of English as a Foreign Language
KW	Keyword		

CHAPTER 1 INTRODUCTION

1.1 Goal of the thesis

Written assessment remains the principal way in which both native speaker (NS) and non-native speaker (NNS) undergraduate students are judged throughout their university studies (Douglas, 2010; Hewings, 1999; Lea, 2004; Leki and Carson, 1994; Lillis, 2001; Lillis and Scott, 2008; Nation, 2008; North, 2005b). Given that success or failure at undergraduate level is likely to have a great impact on the lives and careers of individual students, the ability to write in the preferred ways of the academy (and thereby achieve success in written assessments and ultimately receive a degree) is of considerable importance. However, relatively few large-scale studies have been carried out on assessed undergraduate writing from NS students in the UK, and fewer still have been conducted on NNS student writing, despite the recent rapid growth in numbers of international students in UK universities (UKCISA, 2011, latest figures are 2008/9). The largest group among the NNS undergraduate student body in the UK is Chinese students (UKCISA, 2011) but, again, little research has been carried out on their assessed student writing. The majority of large-scale studies of both Chinese students' writing and NNS student writing in general have been corpus studies concentrating on data sets of *un*assessed, extremely short, argumentative essays collected mainly from non-UK universities (and known as 'learner corpora') (studies of Chinese-only learner corpora include Cross and Papp, 2008; Hyland and Milton, 1997; Wen and Clement, 2003; and studies of writing from groups of NNSs include Granger and Rayson, 1998; Paquot, 2010; Petch-Tyson, 1998). The essays within learner corpora are very different from assessed undergraduate writing since they require neither background reading nor research and consist of prose-only responses to set titles asking for the writer's opinion on a general knowledge issue (learner corpora are discussed further in 2.6). While the results of such studies provide useful insights into features of NNS writing within these particular data sets, this thesis argues that the findings cannot be unquestioningly extended beyond them.

The goal of this thesis is to add to the body of knowledge concerning current undergraduate student writing, through examination of a dataset of Chinese students' assignments

submitted to UK universities between 2000 and 2008. The majority of this dataset is a subset of the British Academic Written English (BAWE) corpus with the addition of a small number of extra assignments from Chinese undergraduates, collected for the purposes of this study. As the sample of texts is moderately large at 279,000 words[1], the methodological approach of Corpus Linguistics can be followed. Adopting a corpus linguistic approach enables the description of the general text characteristics of the dataset, the analysis of one corpus against a comparison or 'reference' corpus, and the extraction of frequently-used language. In line with learner corpus studies, this thesis compares the corpus of Chinese students' assignments with a corpus of texts from first language (L1) English students. However, unlike most researchers of NNS writing (using both corpus and non-corpus methodologies), I do not assume that the NNS writing is less acceptable than that of the NSs simply by virtue of the students' L1, educational and cultural backgrounds. Instead, for both NS and NNS students, I view writing at university level as a gradual process of learning 'meaning making' both within the academy and within particular disciplines (cf. the academic literacies approach of Lea, 2004; Lea and Street, 1998; Lillis, 2001; Street, 1998; and discussed further in Chapter 2). The approach of describing student writing taken in this thesis is thus one of describing *variation* rather than *deficit*, with the recognition that both NS and NNS students are learning to write within the preferred ways of the academy: this approach could be described as 'academic-literacies *informed*' as it is influenced by academic literacies theories yet adopts a different methodology (predominantly Corpus Linguistics). Thus, learning how to write in the preferred ways of a specific situational context (e.g. a particular assignment set by an individual lecturer within their university department at one point in time) is seen as a challenge for both NNS *and* NS university students.

This view of student writing as a contested and situated space is usually researched through the methods of observation, interviews and discourse analysis within an ethnographic approach, rather than through the large-scale text analytic methods of a corpus linguistic approach. Although the main focus in this study is on the final product of the Chinese students' writing (i.e. the assignments as submitted to tutors), background data provided

1 Here and elsewhere the terms 'word' or 'wordform' are used to denote an orthographic sequence of characters separated by spaces. In contrast a 'lexical item' may contain one or more wordforms providing they constitute a single unified meaning.

through questionnaires, interviews and observations of English classes in China assist in giving a sense of the undergraduate students as developing writers and of the changing context of writing.

This first section has set out the goal of the study as the exploration of Chinese students' undergraduate writing in UK universities. The remainder of this chapter discusses Chinese students and their educational background (1.2), describes the challenges posed by assignment-writing at undergraduate level for all students (1.3) and provides an overview of the rest of the thesis (1.4).

1.2 Chinese students

This section first contextualizes the study within the recent rise in numbers of international students in UK universities, justifying the focus on Chinese students as the largest subgroup (as well as one whose assessed writing has been under-researched). The homogeneity of the Chinese students in the sample is examined, and the educational background of students from the largest national grouping, the People's Republic of China (PRC), is explored ('the PRC' is used synonymously with 'China' in this thesis). English Language Teaching (ELT) education within the PRC at primary, secondary and tertiary levels is described in order to understand the experiences and expectations of this student group when they come to write assignments in UK universities.

1.2.1 Numbers of Chinese students in the UK

The number of international students in the UK has been increasing rapidly in recent years and currently stands at over 600,000 per year, estimated to be worth 8.5 billion pounds to the UK economy (British Council, 2010a). Narrowing this student data to undergraduate and postgraduate study in UK Higher Educational Institutions (HEIs) reveals that more than half of all international students in the UK are engaged in tertiary level study. The number of non-UK students in HEIs in 2008/09 was 368,970 (15% of the whole student population in UK HEIs), compared with 325,985 in 2007/08 (14% of the overall student population), an increase of 8% (Table 1.1).

All non-UK domiciled students in HE	Number of students
Postgraduate research	40,275
Postgraduate taught	131,865
Postgraduate other	11,245
First degree	153,355
Other undergraduate	32,230
Total	368,970

Table 1.1 All full and part-time non-UK domiciled students in Higher Education in 2008/9 (table adapted from UK Council for International Student Affairs (UKCISA), 2011)

Within all non-UK domiciled students, the single greatest provider of international students to the UK is the PRC. Moreover, additional regions containing Chinese-speakers are among the top 10 non-EU senders (namely Malaysia, Hong Kong[2], and Taiwan) (Table 1.2).

Since most students from Hong Kong and Taiwan are Chinese-speaking, together with about a quarter of those from Malaysia, this renders the potential number of Chinese-speaking students in UK HEIs (whether speaking Mandarin, Cantonese, or other Chinese dialects) to around 65,000 people. Figures from the British Council's latest country profile report on the PRC gives a slightly higher figure (compared to Table 1.2) of 50,455 students from the PRC in 2008/9, with approximately half this number studying on undergraduate courses and half undertaking postgraduate courses or research (British Council, 2010b: 16). It is unclear whether the figures for students from China entering UK HEIs will continue rising: on the one hand there are very high numbers of people in the PRC currently engaged in English language study (estimated at 330,000 by Bolton, 2008; and 250 million by McArthur, 2008) and household income has rapidly increased in the last decade, enabling families to send their child abroad to study (British Council, 2010b). On the other hand, tighter immigration rules and the growing provision of English-medium universities in the PRC may restrict these numbers (Tallack, 2006). The importance of a continuing influx of Chinese students to both the UK economy and to relations between China and the UK is evidenced by the recent

2 UKCISA provide figures for Hong Kong separately from those for the PRC.

extensive discussions on education at the 5th Annual China-UK Ministerial Summit in November 2010 (British Embassy in Beijing, 2010).

Top 10 non-EU senders	Number of students
China (PRC)	47,035
India	34,065
Nigeria	14,380
Malaysia	12,695
United States of America	14,345
Pakistan	9,610
Hong Kong	9,600
Canada	5,350
Taiwan	5,235
Saudi Arabia	5,205
Total	157,520

Table 1.2 Top 10 non-EU senders of students to UK HEIs in 2008/9 (table adapted from UKCISA, 2011)

In the discussion so far, students from different regions and nations have at times been conflated under the umbrella heading of 'Chinese students'. The next section considers the extent to which this homogeneity can be justified for the purposes of the study.

1.2.2 Commonalities of Chinese students

This research concerns the writing of 'Chinese students', a category in widespread use but constituting a vague denotation as it covers linguistic, ethnic and national groupings. The tendency within English Language Teaching (ELT) in the West to conflate all students speaking any dialect of Chinese, whether from the PRC, Taiwan, Hong Kong, Singapore or Malaysia, under a generic label has recently been widely critiqued (e.g. Clark and Gieve, 2006; Gerbic, 2005; Kennedy, 2002; Pilcher et al., 2006; Sharpling, 2004; Shu, 2006). For example Shu (2006: 139) argues persuasively that, 'when we consider "Chinese students",

we should consider the variety of their national, regional, economic, class and cultural backgrounds as well as age, religion and gender'. However, alongside differences between individuals, there are also important shared characteristics across the whole group such as a literacy based on ideographic characters, a broadly similar language learning methodology resting on Grammar Translation, and a heritage founded in Confucianism. Taken together, these characteristics give some justification for a homogeneous nomenclature, at least for the purposes of analyzing the group's written academic assignments and permitting comparability with previous datasets of written texts ('text' is used throughout this thesis as an alternative to 'assignment'). Each of the three proposed commonalities across Chinese students will now be considered in more detail.

The first shared feature concerns the writing system for the Chinese language. Although there are many varieties of Chinese spoken in East Asia there is a single standard written form, rendering literacy a uniting force (Hu, 2001). While Cantonese, Shanghainese/Wu, Min, Hakka, and so on are often termed 'dialects' of Chinese, only Mandarin (also known as 'Modern Standard Chinese' or 'Putonghua') is consistently deemed a 'language' by Chinese people due to its link with the ideographic writing system (Gao, 2000). A common writing system is an important unifying feature since it enables communication between people with mutually unintelligible dialects, for example the 56 recognized dialects of the PRC (Hu, 2001) and the dialects used beyond the PRC. For previous generations, difficulties in learning the number of Chinese characters required for basic literacy, combined with a lack of access to education, meant that only a small number of educated people were considered 'literate'. A simplification of the writing system and the advent of printing contributed to an increase in the number of people learning to read and write (though learning characters is still a formidable task: Zhang et al., 2005 report that knowledge of 3,500 characters is required to cover 99.5% of modern readings). Today, while people from different regions of the PRC and from other countries may still struggle to communicate with each other orally, a shared written language enables communication and a degree of common culture (although the PRC, Singapore, and Malaysia use versions of the writing system with simplified Chinese characters whereas Hong Kong, Taiwan and Macau use traditional characters).

The second shared characteristic among Chinese students is the method for learning both Chinese and foreign languages, with the contention that the techniques employed in studying the former have strongly influenced all learning. For example, Jin and Cortazzi (2006) argue that the prevalent techniques for learning Chinese characters, namely demonstration, modelling, tracing and copying, influence students' conceptions of how learning takes place more generally. In the same way, Alexander (2001) contends that learning Chinese has shaped the way students learn English:

> Chinese literacy practices seem to encourage Chinese students to approach the learning of English with a similar attention to specific detail and a similar respect for the authority of the teacher (2001: 1).

For Chinese students, then, a focus on extracting and memorizing new words in written texts is applied to foreign language teaching and learning, accounting for the popularity of Grammar Translation (Dzau, 1990; Hu, 2001). Throughout East Asia, Grammar Translation is still the principal method for language teaching and learning, with some influence from the structural linguistic and behaviourist beliefs of Audiolingualism and, more recently, the adoption of techniques from Communicative Language Teaching (CLT) and Task-Based Learning (Littlewood, 2007). Grammar Translation is characterized by a detailed focus on grammatical points and the systematic use of translation in order to achieve the perceived dominant goals of language learning: an ability to read in the target language and the mental discipline of sustained study (Hu, 2001). This attention to detail in the context of the high value accorded to literacy unites all Chinese students.

Outside the PRC and Taiwan, however, the language learning methods are tempered by the more widespread use of English. In Hong Kong, there is an ongoing debate over which of Cantonese, Mandarin or English should be used as the medium of teaching (Hopkins, 2006), which is fuelled largely by discussions as to whether it is beneficial for students to learn through their L1 (for over 90% of Hong Kong people this is Cantonese) (Ping, 2007). In Singapore, English is the official first language and is the medium of instruction in schools. Due to the bilingual policy followed, secondary school students are also required to study

their ‘mother tongue’: Chinese, Malay or Tamil (Singapore Ministry of Education, 2011) (though note that few Singaporean students study in the UK). In Malaysian national schools, the government has experimented with teaching Mathematics and Science through English but will revert back to teaching through Malay from 2012 (Saw and Kesavapany, 2006). For Chinese Malaysians attending independent schools, the language of instruction is Chinese only. Despite national differences in education across the PRC, Taiwan, Hong Kong, Singapore and Malaysia, sufficient shared literacy and foreign language learning methods exist for this to be a uniting feature.

The third commonality across Chinese students from the regions and nations considered here is the shared background of Confucianism: China, Hong Kong, Taiwan, Singapore, Korea, Japan, Thailand and Vietnam are frequently termed ‘Confucian Heritage Cultures’ (CHCs). In these cultures, beliefs surrounding morality and education are grounded in the teachings of Confucius, a teacher and philosopher from two and a half millennia ago (551-479 BC), and have had a great impact on education. The sayings of Confucius are widely known within CHCs, and still underpin attitudes to education. For example, the proverb ‘everything is low, but education is high’ (*wanban jie xiapin weiyou dushu gao*) illustrates the reverence accorded to study, and the maxim ‘being a teacher for only one day entitles one to lifelong respect from the student that befits his father’ (*yiri weishi zhongshen weifu*) denotes the respect and longevity of the teacher-student relationship (originals and translations from Hu, 2001).

In addition to the described commonalities which are applicable to Chinese-speaking people in general, Chinese students undertaking degree courses in the UK are likely to have similar family and educational backgrounds. These include hailing from an urban area (as city schools generally have better English language provision and thereby enable students to meet UK English language requirements), highly valuing the UK education system (and hence wishing to study in the UK rather than in the PRC or other countries), and having a family wealthy enough to support three years of study abroad (one reason for the greater numbers of Chinese people taking UK masters courses rather than undergraduate courses

is that the former entails just one year of fees and living costs). The shared features of Chinese people and of Chinese students in the UK allow a case to be made for treating the group as a whole, at least for the purposes of analyzing student assignments. Linguistic, national and educational differences between Chinese students would be more pressing in a study of strategy-learning or of student motivation. A further, pragmatic reason for considering all Chinese students as a single cohort is that the contextual data in BAWE details only the student's self-proclaimed L1 (for many Chinese students this is simply 'Chinese'), and does not request information on perceived ethnicity of NS or NNS students (discussed further in 4.2.3).

In this thesis the terms 'NS' and 'L1 English' are used synonymously, as are 'NNS' and 'L2 English', though it is recognized that references to 'native' and 'non-native' speakers are contentious (as discussed by, for example, Leung et al., 1997; Römer, 2009). Throughout this study, the student groups are termed 'L1 Chinese', to refer to students who speak any dialect of Chinese and who lived in a Chinese-speaking environment for all or most of their secondary education; and 'L1 English', denoting students whose self-proclaimed first language is English and who lived in the UK for all or most of their secondary schooling; for brevity, these labels are shortened to 'Chinese' and 'English'. While recognizing that a significant minority of Chinese students in this study originate from countries other than the PRC, section 1.2.3 concentrates on the educational background of PRC students as they constitute the largest national group.

1.2.3 Educational background of students from the PRC

Attainment in English language is regarded in China as key for the country's advancement within a global economy. Bolton (2008: 8) points out that English 'has become a marker of middle-class identity, as well as a means for young people to gain an internationally competitive education and employment.' Since 2001, learning English has been a compulsory subject from Grade 3 of primary school to the end of the second undergraduate year at university. Proficiency in English is required both to enter and to graduate from university, no matter what degree discipline is read, and is often a necessity for promotion

across a variety of professions (Hu, 2001). A consideration of Chinese students' experience of English language learning prior to degree study in the UK is helpful as it provides background information on their abilities and expectations concerning academic writing. This section examines the ways in which English is learned in the PRC at primary, secondary and tertiary levels, focusing particularly on the major syllabus component of the Intensive Reading lesson and on the written tasks within exams.

At primary and secondary level in the PRC class sizes for all subjects are large compared to Western classrooms with an average of 50 students, though this is not viewed as problematic since it allows teachers to have a lower teaching load and to spend more time on lesson preparation (Jin and Cortazzi, 1998). However, larger classes in English language mean that it is difficult to focus on spoken communication or engage in discussions around writing, and this tends to promote a transmission method of teaching. Wang and Wen (2002: 228) discuss the lack of attention to productive writing skills, commenting that despite a primary and secondary education in English language of four hours a week for eight years, students receive 'no systematic training in writing'. Kinzley (2011: 202-3) interviewed ten PRC students studying in the UK about their secondary school writing experiences in both Chinese and English classes. He found that in Chinese language classes, writing included 'short paragraphs' (two students) and writing about moral dilemmas (four students), with the required wordcounts given as 2-300 (one student), 800 words (one student) and 2,000 words (one student). Kinzley's interviewees described the aim of the essay-writing as including the improvement of Chinese character knowledge and the development of calligraphy skills. His interview findings for subject classes (where learning took place in Chinese) also indicate that writing is limited to single paragraphs and short answer questions, with the longest essay given as just 600 words; five of the nine students claimed to have done *no* writing at all in their subject classes. Given that so little was written in Chinese, it is unsurprising that the students' writing experience in English language classes is even more limited with students mentioning very short 'essays' describing pictures or giving opinions of just 1-300 words (five students) and diary entries of 2-300 words.

The Intensive Reading programme

The main method of teaching English in China at both secondary and tertiary levels is through programmes of Intensive Reading (Adamson, 2004), described by Gu (2003: 77-8) as 'the single most important source of English input in the Chinese EFL context'. Despite the name, a programme of Intensive Reading is not primarily a reading course but provides a 'core foundation course in EFL in which everything the teacher wants to teach (grammar, vocabulary, reading aloud, etc) is taught through a written text' (Cortazzi and Jin, 1996: 66). An Intensive Reading course involves choral drilling of new words and painstaking analysis of sentence patterns but rarely the discussion of textual content, with the result that students develop a habit of slow reading with word-by-word translation (Dzau, 1990). The typical procedure in the Intensive Reading class is for students to prepare before the lesson by reviewing the provided word list and reading the assigned text. Teachers often use Powerpoint presentations to provide additional information on vocabulary and grammatical structure (Cai, 2011). In class, the teacher leads choral drills and asks individual students to translate sentences which the rest of the class then copy, frequently filling all space around the text with annotations (see Figure 1.1).

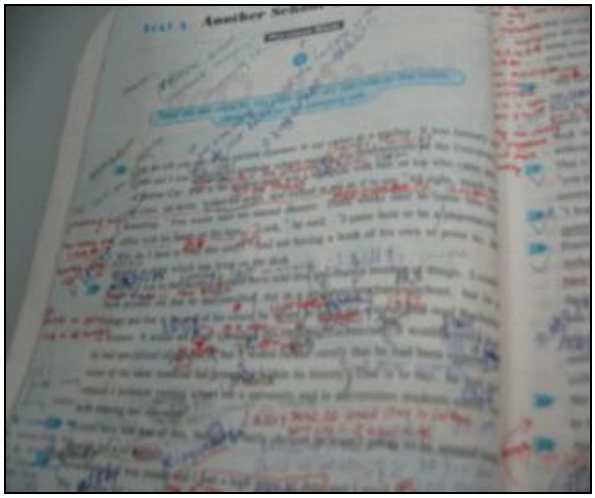


Figure 1.1 Example of student's annotated textbook[3]

The text is used to draw out particular grammatical points which are discussed in detail in Chinese. This focus on repeating vocabulary items and analyzing grammar leaves little time

3 I am grateful to Guozhi Cai for permission to use the photograph in Figure 1.1.

for freer discussion (Cheng, 2000). Instead, Chinese teachers 'tend to stick to the textbook, which is often the same… throughout practically the whole country' with great emphasis placed on learning grammar and vocabulary as this is what is tested in exams (Boyle, 2000: 153). Textbooks are prescribed by the state and are heavily exam-oriented, containing vocabulary lists, cloze exercises, and unconnected sentences exemplifying single grammatical points. Even textbooks which claim to be adopting a Communicative Language Teaching (CLT) approach with its focus on communication maintain the 'centrality of grammar' and 'transmission modes of learning' (Tomlinson, 2005: 7). This reliance on memorization techniques in the Intensive Reading procedure should not, however, be seen as a 'mechanical rote learning experience' but rather a way of learning which involves 'careful thinking, appreciation of rhetoric, meaning making and understanding' (Gu and Brooks, 2008: 347).

An advantage of the Intensive Reading lesson is that it is unthreatening to the NNS teacher, since preparation and the following of a prescribed reading lesson format means the ensuing lesson is highly predictable (Adamson, 2004), in sharp contrast to the advanced language level and sociolinguistic competence required by teachers following a CLT approach (Hu, 2005). CLT regards the acquisition of communicative competence (Canale and Swain, 1980; Hymes, 1972) to be the inherent goal of language learning with a resultant focus on learner-centred interaction, reduction of the use of L1, and preference for authentic materials, situations, activities and tasks (Hu, 2001; Richards and Rodgers, 2001). Hu (2001) points out, however, that CLT in China struggles to compete with the longevity and teacher-centredness of Grammar Translation and that many Chinese teachers adopt an 'eclectic' approach (Jin and Cortazzi, 2002: 59), employing techniques from a variety of methods and approaches. This eclecticism was also apparent in my observations of ELT classrooms at primary and secondary level in Beijing in 2008 and 2010. It is likely that teachers are motivated to retain the traditional method of Grammar Translation with its focus on accuracy and sentence-level grammar due to the importance of university entrance exams in the PRC (and in East Asia more widely). The ensuing 'washback' effect, wherein exams strongly

influence both *what* is learned in class and *how* this is learned, means that change can be difficult to introduce into the English language classroom (Qi, 2004, in Kinzley, 2011).

Preparing for University

At secondary level, children face intense pressure to succeed from teachers, parents and themselves, in the form of extensive homework, and additional 'cram school' preparation for the 'Gaokao' or university entrance exams (Dongping, 2006). The One Child Policy has increased this stress as there may be six adults (parents and four grandparents) expecting a single child to gain a place at university (Gu and Schweisfurth, 2006), though a consequence of having only one child is a reduction in the financial burden. The older generations' memory of the dearth of educational opportunities during the Cultural Revolution heightens the pressure placed on current school students. Increasing numbers of students in the PRC wish to attend university yet there has not been a corresponding rise in the number of places within HE, resulting in fierce competition. In 2010 nearly 10 million high-school students sat the Gaokao for an estimated 6.6 million undergraduate places in the PRC (Liang, 2010).

English language now has equal status with Chinese and Mathematics in the Gaokao, thus serving a gatekeeping function of restricting the number of students entering university. In 2000 the English language syllabus was changed to become more communicative, the National Matriculation English Test (NMET) was adopted and new textbooks were developed. However, as the test format still primarily comprises multiple choice items, the result has not been a sea change in the methods used in ELT (Qi, 2004, in Kinzley, 2011). Qi found that instead of encouraging a process approach towards writing, the NMET encouraged teachers to emphasize the multiple choice sections of the test as this proved the most efficient way to improve students' scores and also as teachers did not have time to provide feedback on longer stretches of writing. In the Chinese section of the Gaokao, the writing tasks also provide some insight into valued academic writing practices. These tasks include summarizing a text of 15,000 Chinese characters to one of 800, re-ordering paragraphs into a more logical sequence, and adding missing paragraphs to a gapped text (Kirkpatrick, 2004). Great attention is paid to the structure of the Chinese language, while the ability to write critically is accorded less importance. As with the English exams, this skills-

based approach gives students the tools to manipulate language but provides little practice in extended writing.

Once at university in the PRC, English language classes for both English majors (students specializing in English language and literature) and for the majority group of non-English majors (students taking compulsory classes for the first two years of study) continue the methods of secondary level teaching with an emphasis on the 'five skills' of listening, speaking, reading, writing and translation (Jin and Cortazzi, 2006). Students intensively study short texts, as in secondary school, and make notes from their tutor's Powerpoint slides. Content continues to be exam-oriented and heavily centralized, with lecturers having little autonomy in deciding course content (Wang, 2003). Again, class sizes are generally larger than in the UK, often with 50 students in a 'discussion' class. Writing courses in English at PRC universities are either 'translation courses' or 'composition courses', though it seems that neither teachers nor students are willing to give a great deal of time to the development of extended writing in English (Dzau, 1990; Qi, 2004 in Kinzley, 2011).

Study abroad

The alternative to the intense competition surrounding university entrance in the PRC is to compete, and pay, for a degree course in an overseas university. For students taking this route, an English language test is imperative. The majority of Chinese students entering UK universities sit the International English Language Testing System (IELTS) as a prerequisite for entry to a degree course with increasing numbers of Chinese students taking IELTS each year (Cambridge ESOL, 2011) ('ToEFL' is the Test of English as a Foreign Language and is the English language proficiency test used most commonly in the US). Fox and Curtis (2010) conducted interviews on the topic of IELTS practice with 15 students from the PRC and Hong Kong, finding that test preparation is based around students studying practice tests every day for weeks or months beforehand. Similarly, Cheng (2000: 444) comments that in the PRC, 'ToEFL takers spend hundreds of hours doing simulated tests to develop test-taking strategies rather than improving their real language skills'. Notwithstanding the drawbacks of this mode of practising for IELTS and ToEFL, the actual written essays required in the test are poor preparation for academic writing in English as they are

extremely short, divorced from reading, and encourage students to provide opinions drawn from their life experiences (cf. discussion in Moore and Morton, 2005, on IELTS tasks; and further discussion in 2.6.2). In the 60-minute Writing section of IELTS, the suggested minimum word counts are 150 for Task 1 and 250 for Task 2. For ToEFL, the paper is even shorter at just 50 minutes with no specified minimum word count.

Shen (2005) proposes a range of reasons behind the increase in numbers of Chinese undergraduate and postgraduates studying abroad, discussing these in terms of 'push' and 'pull' factors. Dominant 'push' factors are the intense competition for University entrance exams in China and the longer time spent gaining a degree: four years at undergraduate level and three years for masters level compared with the UK norm of three years and one year respectively. 'Pull' factors include the 'prestige of foreign degrees and advancement in English and foreign languages' (Shen, 2005: 430) as well as a 'general recognition of the cultural benefits of studying and living in another country' (British Council, 2010b: 12), all of which are seen to improve the chances of gaining a good job on graduation. In many cases it is not the Chinese student's sole decision as to which country or even which course of study is chosen, with parents' views being paramount (Gieve and Clark, 2005). Chinese students bear the expectations of their parents on their shoulders to a far greater extent than is the case for UK students; children feel an obligation to repay parents for the debt incurred in raising and educating them (Wang, 2003) and are concerned with gaining good grades to show their parents they are working hard (for example, several Chinese students contributing to the BAWE project requested that certificates be provided to prove they have contributed to a research project which accepts only proficient student writing).

Although there are many positive factors encouraging Chinese students to undertake degree courses in the UK, study abroad also has its challenges in accommodating to different teaching methods and to writing in a foreign language. The next section examines challenges of assignment-writing for both NS and NNS students within UK HE.

1.3 Challenges in assignment-writing

Academic success in UK HEIs rests on an ability to write well (e.g. Douglas, 2010; Hewings, 1999; Leki and Carson, 1994; Lillis and Scott, 2008; Nation, 2008; North, 2003, 2005a). Writing is important because it is the main way in which students demonstrate that they have learned and understood the discipline content, it illustrates an ability to employ the genres which are valued within the academy, and it enhances students' capacity to think and thereby increases their cognitive development (Hewings, 1999). The significance attached to written assessment in the UK has led Lillis and Scott (2008: 9) to describe writing as a '"high stakes" activity in university education' such that 'if there are "problems" with writing, then the student is likely to fail' (Chambers and Northedge, 1997, similarly distinguish between 'high stakes' and 'low stakes' writing on the basis of whether or not it is assessed). This section first considers general challenges for all students, whether NS or NNS, then focuses on additional challenges faced by NNS students and by Chinese students in particular.

1.3.1 Challenges for all students

All undergraduate students in UK Higher Education currently face a number of challenges in writing these high-stakes assignments (Harwood and Hadley, 2004; Lea and Stierer, 2000a). Difficulties include tutors' lack of articulation as to exactly what they require (Crème and Lea, 2003; Lea and Street, 1998; Lillis, 1997), tutor and students' varied ideas of what a particular assignment entails (Durkin and Main, 2002; Elton, 2010; Lea, 2004) and different perceptions of what constitutes 'good writing' (Lea and Stierer, 2000b; Lea and Street, 1998; Lillis and Turner, 2001).

The major strategic aim of assignment-writing is to display disciplinary knowledge in an appropriate form in order to gain marks and ultimately a university degree (Kaldor and Rochecouste, 2002). Framing this knowledge, however, is difficult since undergraduates are effectively required to write for a dual audience: the assignment rubric may necessitate writing for an audience with little knowledge of the discipline (the imagined non-specialized reader of the essay/press release/case study), while the purpose of assessment involves writing for display purposes to an audience with a high level of disciplinary knowledge (the

discipline lecturer/assessor) (cf. Lillis', 2003, ethnographic account of students' struggles with this dual audience). For Tang (2009: 15) the dualism is framed through students' efforts to engage with the discipline; she draws on Bakhtin's (1981) notion of dialogism in her discussion of students attempting to both participate in the '"dialogue" of a wider discourse community' and answer an assignment 'question' for a specific tutor/reader. The notion of an 'assignment' is a very different form of writing to other academic genres since, as Abasi and Akbari (2008: 25-6) point out, the label 'implies a literate task imposed by a person in authority, which by itself serves to construct a teacher-student relationship'.

Many researchers have emphasized how university students have to learn to write in ways prescribed by their discipline in order to have their voices recognized (Bazerman, 2001; Harwood and Hadley, 2004; Hewings, 1999; Hyland, 2008b; Lillis, 1999, 2001; North, 2005b; Prior, 1998; Rai, 2008). Corpus studies such as Hyland (2002b, 2004, 2008b) and ethnographic studies such as Prior (1998) have illustrated the extent to which academic writing varies between disciplines. Indeed, Harwood et al. (2004: 366) suggest this variation extends beyond whole disciplines to practices in academic writing which differ 'from department to department, and even from lecturer to lecturer' (see also discussion of disciplinary difference at undergraduate level in 2.2). Despite these demands, classes in English for Academic Purposes (EAP) frequently consist of students from a wide range of disciplinary areas. Such varied EAP classes rarely practise the writing required within, or even distinguish between, each student's disciplinary area. Moreover, these classes are frequently only offered to NNS students, whereas both NS and NNS students have to learn to write in ways valued by both institution and discipline in order to succeed at university.

As well as the long-standing challenge of working out the 'rules of the academic achievement game' (Newman, 2001: 470) and of learning to write within a discipline, students are faced with more current challenges. In recent decades, UK Higher Education has altered from a 'conventional single route initiating a cohort of students into the practices of their discipline' (North, 2005a: 517) to a system of increasingly flexible modules encouraging inter-disciplinary degrees. Since students at undergraduate level are likely to

take courses from more than one discipline, and given that there is often a gap between lecturers' and students' understandings of assessment criteria (Durkin and Main, 2002), students must adapt more quickly than ever to writing within discipline areas which are new to them.

Within each module in each discipline, students must also comprehend the genre expected of them in written assessments. A recent, additional challenge in UK HE is the 'unprecedented amount of innovation in assessment' (Gibbs, 2006: 20) giving rise to an array of innovative assignment genres (Ganobcsik-Williams, 2004; Leedham, 2009; Nesi and Gardner, 2006). The prose essay, case study and laboratory report used to provide the mainstay of undergraduate writing, but now students may also be expected to produce blogs, letters and e-posters, presenting difficulties for all undergraduate students in the UK, and adding significantly to the existing pressure of producing extended writing for L2 English students (Leedham, 2009) (see list of BAWE genres in Appendix C). Despite these new genres, any difficulties in comprehending what is required of the student in completing a set task still appears to be viewed as the student's 'problem' rather than a lack of support from lecturers (Gourlay, 2009).

1.3.2 Additional challenges for Chinese students

All NNS students face significant time pressures through needing longer to read and write academic texts (Leki and Carson, 1994; Luxon and Robinson, 2006; Mauranen, 1994; Whitley, 2007). Whitley (2007: 9) comments that Chinese students felt the workload was 'heavier' than they expected and Luxon and Robinson (2006) suggest that typically Chinese students take two or three times longer than British students to do the required reading for a course. Reasons for Chinese students' difficulty in reading include a lack of familiarity with Romanized script, and perhaps also the fact that alphabetic scripts provide information in an elongated way when compared with the 'compact ideograms of Chinese' (Swan and Smith, 2001: 313). Similarly, Bassetti (2009: 772) suggests that L2 reading 'is not simply inefficient reading, it is qualitatively different from native readers' reading'. It is also likely that the Intensive Reading programme of the PRC with its focus on word by word translation is

responsible for slower reading rates, since students are not accustomed to dealing with the ambiguity of unknown lexis and therefore tend to consult a dictionary for each unfamiliar item.

For Chinese students accustomed to the tightly-controlled didactic teaching exemplified by Intensive Reading lessons and Grammar Translation method, moving to the more student-centred classes of UK universities is a challenging experience (Cross and Hitchcock, 2007; Kinzley, 2011; Tian, 2008). Cross and Hitchcock (2007) explored Chinese students' attitudes at the University of Portsmouth and found that they were very aware of differences in the roles of learners and tutors between the UK and China. In the PRC, the teacher tells the student what they need to know and the student's role is to absorb knowledge 'like a memory stick'; in contrast, in the UK the relationship between student and teacher is 'like traveler and guide' (p.9, quotations from student respondents in Cross and Hitchcock's study). Similarly, Tian (2008) interviewed Chinese postgraduate students in the UK and found that they experienced challenges in comprehending the roles of teachers and students, and also in learning to write in a more critical way than in their previous undergraduate study in the PRC.

The 'discourse universe' encountered by first year students at university may be 'something of a culture shock, because the university institution cultivates its own distinctive discourse types which are not used in the outside world' (Mauranen, 1994: 1). For the majority of Chinese students in this study, writing undergraduate assignments in their UK university represents their first real experience of any tertiary-level writing, whether in Chinese or in English (most English students are also, of course, new to degree-level writing). Once studying in a UK university, Chinese students may find the assessment very different from that experienced previously with a variety of genres of assignments expected, in comparison with the dominance of traditional written essays in China. For example, the large cohort of Chinese students with a secondary education from the People's Republic of China are more accustomed to the limited genre of 'traditional written exams' relying on the memorization of knowledge (Cross and Hitchcock, 2007: 9) than the plethora of different genres encountered

in UK universities. The use of oral presentations and reflective writing (i.e. writing in which the student retrospectively analyzes the process of carrying out a task) are singled out by Cross and Hitchcock as likely to be particularly unfamiliar for Chinese students. Tian (2008: 146) similarly comments on the 'fact-oriented exams' which are the main assessment in China compared to assessment in the UK with its greater emphasis on critical thinking and on learning strategies.

This section has explored some of the challenges of writing assignments in UK HE. In addition to the lack of clarity and consistency of assignment rubric, learning how to write in discipline and lecturer-approved ways, and the increase in interdisciplinary degrees, students face a rapid expansion in the number of assignment genres. While both English and Chinese students must negotiate these challenges, Chinese students face additional hurdles such as learning to read quickly in their L2 rather than in the intensive manner encouraged in the Chinese education system, and producing longer pieces of writing than they are accustomed to. This study examines Chinese students' writing to explore similarities and differences to English students' writing and, through investigation of particular features, gain some insight into how the challenges of writing at undergraduate level are met. While the focus of the study is on Chinese students, the findings may also have implications for all NNS and for NS students.

1.4 Organization of the thesis

This final section provides an overview of the organization of the rest of this thesis.

Chapter 2 turns to features of Chinese students' writing within the research literature in order to motivate the investigatory areas within the current study. Four review questions are first posed of the available literature in this area, concerning (i) characteristics of English students' undergraduate writing, (ii) characteristics of Chinese students' undergraduate writing, (iii) how these features may vary across year groups, and (iv) the effect of writing in different disciplines. Chapter 2 also discusses the research questions (RQs) for this study, placing these in the context of the reviewed literature. RQ1 concerns the characteristics of

Chinese students' writing at undergraduate level, in comparison with English students' writing, and the extent to which these characteristics reflect those identified in the research literature. RQ2 concerns the extent to which these characteristics vary at different stages of the undergraduate journey. Finally, RQ3 examines the effect on the identified characteristics of writing within different disciplines.

Chapter 3 concerns the linguistic phenomenon of lexical chunks, framed as the selected tool for exploring the student assignments. Throughout this thesis, 'lexical chunk' or simply 'chunk' is used to refer to any linguistic item which may be classed as a unit, whether through its frequency, internal coherence, or other linguistic features; 'phraseology' is similarly used as a superordinate category encompassing all types of chunk. The use of further terms is explained in Chapter 3. The chapter reviews the wide range of theories relating to lexical chunks, focusing on Hoey's (2005) lexical priming as a broader theory of language learning. An examination of the characteristics often ascribed to chunks, methods of identification, and the resulting multitude of terms used for phraseological elements is followed by a discussion of the structural and functional classification of chunks, following Biber et al. (1999) and Hyland (2008a,b).

Chapter 4 focuses on the data and research methods used within the study. It first traces the process carried out to compile the datasets used in the study; these comprise data taken from the BAWE corpus as well as additionally-collected assignments from a range of UK universities. The chapter then describes the corpus linguistic procedures used to conduct an initial profile of the text characteristics of each corpus, to extract keywords and key chunks (that is, items which are used significantly more than would be expected in comparison with a reference corpus), and to identify and analyze lexical chunks in the corpora.

Chapters 5, 6 and 7 provide the main corpus findings. **Chapter 5** focuses on the overall findings, seeking to answer RQ1 concerning characteristics of Chinese students' writing overall and how these reflect those identified in the research literature. Following the extraction of keywords and key chunks from the data, four overall differences are found

between Chinese and English students' academic writing, three of which echo observations from the literature (use of first person plural, (limited) use of informal language and reliance on particular connectors).

Chapter 6 considers the extent to which the identified characteristics vary across year groups, again using keyword analysis of the corpora (RQ2). The focus is then narrowed to four-word chunks, and these are categorized structurally and functionally as this enables the data to be grouped and compared. The chapter suggests that the first three differences discussed in Chapter 5 between the student groups are most prevalent early on in undergraduate writing and less apparent in the year 3 texts, perhaps as both groups conform to the expected academic norms.

Chapter 7 focuses on disciplinary differences (RQ3). The three disciplines most comprehensively represented in the Chinese corpus are Biology, Economics, and Engineering, and these are compared with reference corpora from the same disciplines in the English corpus. An investigation of the use of visuals and lists in pairs of texts (answering the same assignment title) from Biology, Economics and Engineering illustrates the multimodal nature of the assignments and the range of acceptability permitted.

Chapter 8 concludes the thesis, drawing together the findings from Chapters 5 to 8. The main contributions of the study are discussed; these include relating the findings identified in Chapter 5 - 8 from corpora of undergraduate assignments to those observed in the learner corpus literature discussed in Chapter 2. Finally, some limitations of the study are outlined, and suggestions are made as to how these could be overcome and the findings extended in future studies.

1.5 Chapter summary

Chapter 1 began by setting out the overall goal of the study: to research features of Chinese students' undergraduate writing in English in UK universities. The chapter has explored the motivation behind this study of Chinese students' assignment writing, detailing the size of

this group within UK HE and the lack of prior research into their undergraduate writing since the majority of studies focus on datasets of short, argumentative essays. The inclusion of Chinese-speaking students from the PRC, Hong Kong, Taiwan, and to a lesser extent Malaysia, and Singapore, was justified on the grounds of commonalities across these groups and as a pragmatic decision for the exploration of student writing. A discussion of PRC students' education, particularly in English language, and their experiences of writing in both Chinese and English prior to UK study, provided background information on the cohort. Instead of the common approach of viewing student writing as in some way deficient, with NNS writing being particularly problematic, this study draws on research in academic literacies and views student writing for both NSs and NNSs as a contested space in which participants are learning to write in the preferred ways of both the academy and the discipline. To this end, while the main aim of the study is to explore Chinese students' writing at undergraduate level, findings may also impact on student writing for all NNS and NS students.

Chapter 2 situates the goal of the study in the context of existing research from both the limited number of studies on Chinese undergraduate writing and the greater range of studies on texts from mixed L1 groups of NNS undergraduate students writing short essays. This literature is explored through the lens of four overarching review questions, and at the end of the chapter the answers to these initial review questions motivate the three research questions of the study.

CHAPTER 2 CHINESE STUDENTS' WRITING: THE LITERATURE

Chapter 1 identified Chinese students as the largest NNS group within UK HE, and justified the inclusion of texts from a range of national and dialectal groups in the study. The educational background of all Chinese students was discussed, particularly their experience of writing in English, as this impacts on their studies in the UK. Current challenges for all undergraduate students were also reported in order to situate the task of assignment-writing for Chinese students within the context of UK HE. This chapter surveys findings from the research literature to establish what is already known about Chinese students' academic writing, and to determine what this study can offer to the field. The deficit approach of much of the literature on student writing is also discussed and is contrasted with the descriptive approach taken in this study. Finally, this chapter discusses the research questions which provide the focus of this study.

2.1 Initial review questions

Section 1.1 gave the goal of this thesis: to add to the body of knowledge concerning current undergraduate student writing, through examination of a dataset of Chinese students' assignments submitted to UK universities between 2000 and 2008. The initial review questions in this section arose from this goal and framed my reading of the available literature on Chinese (and to a lesser extent, English) students' writing. While the thesis is primarily concerned with Chinese students' writing, as this is compared with a larger corpus of English students' writing, a preliminary review question focusing on English students (question 1) is first considered to set out what is known about their undergraduate writing before turning to the writing of Chinese students. The overarching review questions numbered 2 – 4 are tightly-framed within the relatively narrow research field of Chinese students' undergraduate writing in UK universities. However, since there have been few empirical studies in this precise area, a broader array of literature has been surveyed, covering Chinese students' writing at foundation or pre-university level, undergraduate and

postgraduate levels; research both in the UK, and in the PRC and Hong Kong; and including the large area of research within learner corpora worldwide. The category of learner corpus literature features studies of Chinese students only, as well as more general studies of NNS writing. The findings from the broader literature beyond undergraduate level are discussed in terms of how they may relate to Chinese students' undergraduate writing.

The four broad review questions are as follows:

Q 1: What characteristics of English students' undergraduate writing in UK universities have been identified and discussed in the research literature?

Q 2: What characteristics of Chinese students' undergraduate writing in UK universities have been identified and discussed in the research literature?

Q 3: According to the research literature, how do identified characteristics of Chinese students' undergraduate writing in UK universities vary over the period of undergraduate study?

Q 4: According to the research literature, how are the identified characteristics of Chinese students' undergraduate writing in UK universities affected by writing conventions in different disciplines?

These review questions were then divided into the following subquestions in order to more precisely focus my reading:

Q 1: What characteristics of English students' undergraduate writing in UK universities have been identified and discussed in the research literature?

1.a. What empirical studies are available on English students' undergraduate writing in UK universities?

1.b. What are the major findings from these studies?

Q 2: What characteristics of Chinese students' undergraduate writing in UK universities have been identified and discussed in the research literature?

2.a. What empirical studies are available on Chinese student writing in English at all levels of academic study?

2.b. What are the major findings from these studies?

2.c. To what extent has this research area been covered?

Q 3: According to the research literature, how do the identified characteristics of Chinese students' undergraduate writing in UK universities vary over the period of undergraduate study?

3.a. What is known about variation in academic writing by English students at different levels of undergraduate study?

3.b. What literature is available which traces variation in academic writing by Chinese students at different levels of undergraduate study?

3.c. What are the major findings from these studies?

3.d. To what extent has this research area been covered?

Q 4: According to the research literature, how are the identified characteristics of Chinese students' undergraduate writing in UK universities affected by writing conventions in different disciplines?

4.a. How can the concept of a 'discipline' best be defined and operationalized for the purposes of this study?

4.b. What is known about how writing varies within different academic disciplines by English students at undergraduate level?

4.c. What does the literature tell us about Chinese students' writing in different disciplines?

4.d. To what extent has this research area been covered?

This section has set out the overall review questions put to the literature on Chinese students' writing. The following four sections survey the literature pertaining to each review question and corresponding set of subquestions in turn: characteristics of English students'

writing (2.2), characteristics of Chinese students' writing (2.3), variation over time (2.4), and disciplinarity (2.5).

2.2 English students' undergraduate writing

Review Question 1: What characteristics of English students' undergraduate writing in UK universities have been identified and discussed in the research literature?

Since English students represent the majority group of undergraduate students in UK universities, this initial review question is intended to provide the context for the later exploration of Chinese students' writing. As discussed in Chapter 1, the thesis takes a descriptive, academic literacies-informed stance towards different varieties of writing, rather than a normative stance in which one variety is viewed as superior to the rest. Thus, the aim of this section is to set out what is known about undergraduate writing from studies of the largest group of such writers in UK universities.

Review question 1.a. What empirical studies are available on English students' undergraduate writing in UK universities?

The number of corpus studies of undergraduate student writing in UK universities was until recently very low due to the lack of widely-available corpora, and was limited to individuals' collections of student writing within one discipline and a single institution (e.g. Hewings' study of Geography undergraduates at the University of Birmingham, 1999, 2004; and North's work on Open University students in the discipline of History of Science, 2003, 2005a). In contrast, aspects of postgraduate writing in the UK and professional academic writing have been more widely studied across a range of disciplines and genres (e.g. Charles, 2003, 2006; Groom, 2005, 2007; Harwood, 2003; 2005; Hyland and Tse, 2005; Oakey, 2002, 2009; Pecorari, 2009; Samraj, 2002, 2005; Thompson, 2001, 2005; Yeung, 2007).

One reason for the lack of research in undergraduate writing is that, unlike postgraduate theses or published research articles, undergraduate assignments are largely outside the

public domain. Undergraduate work is not usually stored in a central repository within an institution (in the way that masters or PhD theses or published articles may be) and individual permissions must be sought from a large number of tutors and students in order to gain access to sufficient data for a corpus study. In terms of funded projects, Nesi et al. (2005: 1) comment that previous large academic corpora have focused on 'professional and semi-professional writing in the public domain' rather than student writing, citing the case of the ToEFL 2000 Spoken and Written Academic Language Corpus (Biber et al., 2002) which includes seminar transcriptions, textbooks, and written course information, but lacks any samples of student writing.

The British Academic Written English (BAWE) corpus (Nesi, 2008a; 2011) is thus the first widely-available corpus containing texts from undergraduate students across a range of disciplines and from several UK universities. All writing in BAWE is deemed 'proficient' student writing, defined as graded assignments receiving the UK Honours degree classifications of Upper Second (or 'merit', with a mark of 60 - 69%) or First (termed 'distinction', achieving 70% plus). Indeed, Nesi et al. 2004: 444, assert that the University of Warwick (a participating university) 'is a multicultural, multilingual environment, and in their departments students are assessed on merit, without regard for their language background' (cf. comments by Sharpling, 2010, on the proficiency of all BAWE writers' texts whether NS or NNS writers). Note that research into BAWE assignments thus includes proficient NNS writing but since English L1 texts form the majority, studies using BAWE are considered in this section.

The release of the BAWE corpus has led to a number of recent studies (e.g. Bruce, 2010; Gardner, 2008; Gardner and Holmes, 2009; Jung, 2011; McKenny, 2005; Nathan, 2010; Nesi and Gardner, forthcoming, 2011; Thompson, 2009). These studies focus on a range of aspects of undergraduate writing and employ a variety of frameworks. For example, Bruce (2010) compares textual and discoursal resources within essays in the disciplines of English and Sociology; Gardner (2008) discusses Halliday's (1994) verbal and mental processes within academic disciplines ('verbal process' and 'mental process' are terms used by

Halliday for verbs connected with speech and cognition respectively); and Thompson (2009) focuses on shared disciplinary norms and individual traits across disciplines. Genre has been the focus of several studies with some researchers focusing on a single genre and discipline; for example Jung (2011) examines the genre of laboratory reports in Engineering and Nathan (2010) focuses on pedagogical Business case reports. Other studies compare genres within different disciplines. For example, Gardner and Holmes (2009) examine the use of headings in assignments from different genres and disciplines, and Nesi and Gardner (forthcoming, 2011) examine reflective writing across disciplines in the BAWE corpus. In a study employing a wide range of different methodologies, Gardner (2008) compares the disciplines of History and Engineering through ethnographic, multidimensional, corpus linguistic and systemic functional approaches to genre analysis.

Review question 1.b. What are the major findings from these studies?

Many of the studies listed in answer to 1.a. concentrate on writing within a single discipline (e.g. Hewings, 1999; Jung, 2011; North, 2003; Nathan, 2010); or focus their study on a comparison of disciplines (e.g. Bruce, 2010; Gardner, 2008; Hewings and North, 2006; Thompson, 2009). The overriding finding of studies comparing disciplines is that distinct differences exist between the assignments produced by students in different disciplines (e.g. differences in the genres required are noted by Bruce, 2010; Gardner, 2008; Thompson, 2009; difference in first person pronoun usage (i.e. the use of *we* and *I*) is discussed by Thompson, 2009; and different uses of Halliday's, 1994, sentence-initial Theme were found by Hewings and North, 2006). Since disciplinarity is the focus of review question 4 in 2.5, a sample of these studies are discussed in this later section in relation to English students' writing across subject areas at undergraduate level. Other studies have commented on the extensive influence of genre within undergraduate writing (e.g. Gardner, 2008), adding support to Moore and Morton's (2005) assertion of the diversity of writing required at undergraduate level in the UK. Genre in undergraduate assignments is also discussed in 2.5 within disciplinarity.

Few writers have compared English undergraduate texts with postgraduate or expert texts (though comparison with expert texts is common in studies of L2 writing). One study which

does compare the two is Thompson's (2009) comparison of the undergraduate Engineering and History assignments in the BAWE corpus with Hyland's (2008a) corpus of postgraduate and expert writing (though note that the postgraduate portion of Hyland's corpus is written by L2 Cantonese speakers). Thompson found that '[t]he most striking difference' (p.65) is the frequency of the pattern '*it* + be + ADJ + *to/that*' (e.g. *it is important to, it is clear that*). Thompson comments that this pattern enables the writer to make an evaluation of a proposition or process. Excepting these comments on the two disciplines overall, most of Thompson's study concerns variation across year groups, and variation across the two disciplines, and is discussed further in consideration of these aspects of L1 writing in 2.4 and 2.5 respectively.

2.3 Characteristics of Chinese students' writing

Review Question 2: What characteristics of Chinese students' undergraduate writing in UK universities have been identified and discussed in the research literature?

Review question 2.a. What empirical studies are available on Chinese student writing in English at all levels of academic study?

In this section, the available studies are classified according to the sources of the data, beginning with university assignments and moving outwards to learner corpora which, while prevalent in NNS writing studies, are somewhat further removed from the assessed student writing carried out within university disciplines.

Studies of Chinese university students in the UK

Until the last few years, corpus research was extremely scarce in the specific area of undergraduate assignments produced by Chinese students in UK universities. Since Chinese students are the largest NNS group within UK HE (as discussed in 1.2.1), assignments from this group are prevalent in the BAWE corpus. The availability of this corpus has thus given rise to a small number of studies focusing on Chinese students' writing (e.g. Chen, 2009; Chen and Baker, 2010; Lee and Chen, 2009; Li, 2010). Chen

(2009; also Chen and Baker, 2010) investigates the use of lexical chunks by Chinese writers in the BAWE corpus, English writers in BAWE, and professional academic writers. Notably, Chen explores undergraduate and masters level writing by Chinese students within a single corpus, compiled on the basis of one text per student, and does not distinguish the texts by level of study, discipline or genre. The issue of level of study is particularly pertinent since postgraduate writing may have had input from supervisors. Lee and Chen (2009) look solely at undergraduate writing from BAWE, examining five 'overused' items from an initial keyword analysis. Li's (2010) study is limited to analysis of metadiscourse within undergraduate writing by Chinese students. All of these studies take a broadly deficit approach, viewing differences between Chinese and NS student or expert academic writing as 'problems' to be remedied through English for Academic Purposes (EAP) classes and materials. For example, Chen (2009: 44) states that her study intends to 'reveal the potential problems in second language learners' written performance from the viewpoint of frequency-defined phraseology by making comparisons with native writing'.

In addition to studies from the BAWE corpus, a number of case studies exist of Chinese students' writing (e.g. Kinzley, 2011; Li, 2009; Li and Schmitt, 2009; O'Connell and Jin, 2001). Kinzley (2011) conducted a case study of 99 essays from 11 PRC students on the same course at the same university, using discourse analysis to establish the extent to which the practices of academic writing advocated on an EAP course were adopted by the students. The number of individual texts covered in Kinzley's study is unusually large for a discourse rather than corpus analytic study, though as these are from one discipline within a single institution the findings may not be generalizable. Most non-corpus studies have used much smaller datasets, making it hard to generalize further from findings as variables such as the particular course and institution, and the particular individuals play a larger part. For example, a case study by Li (2009) and Li and Schmitt (2009) focuses on lexical chunks in the writing of a single masters student from China, detailing changes in the student's use of chunks and examining her view of these changes.

Further non-corpus research has focused on student attitudes towards writing and study in UK universities (e.g. Cross and Hitchcock, 2007; Durkin, 2010; Gan et al., 2004; Tian, 2008) and use of writing strategies (e.g. Gu, 2003; Lei, 2009; Mu and Carrington, 2007; Wang and Wen, 2002). Since these studies do not primarily concern actual texts, they are not considered further in this review.

Studies of Chinese university students in the PRC and Hong Kong

Studies of Chinese students' writing within universities in the PRC and Hong Kong may also contain relevant findings for Chinese undergraduates in the UK. For example, Flowerdew (2003) analyzes year 2 and 3 undergraduate technical reports from Hong Kong students in one institution, comparing these to professional reports. Hyland (2008a) explores postgraduate students' writing in Hong Kong, though he concentrates on disciplinarity and student level (masters and PhD) and does not discuss the fact that the writers are all Hong Kong Chinese.

Learner corpora

While the number of studies of assessed university writing has increased in the last few years, the majority of research into the academic writing of Chinese students has been carried out on short pieces of argumentative writing in collections of 'learner corpora' (e.g. Chuang and Nesi, 2006; Cross and Papp, 2008; Guo, 2006; Hyland and Milton, 1997; Lu, 2002; Milton, 2001; Milton and Hyland, 1999; Wen et al., 2003; Yang, 2005). A number of Chinese-only learner corpora exist which are compiled from university students in the PRC. From the 107 learner corpora[4] listed on the comprehensive Université Catholique de Louvain website (http://www.uclouvain.be/en-cecl-lcWorld.html), I identified 12 written corpora of Chinese students' texts, of which just three contain written texts from PRC students. The first of these is the written portion of the Bilingual Corpus of Chinese English learners (BICCEL), containing of 0.5 million tokens of timed, in-class essays from third and fourth year students in 30 universities across the PRC. The second is the Written English Corpus of Chinese Learners (WECCL 2.0), containing 1.25 million tokens of timed and untimed argumentative and narrative essays written by English majors (Qiufang et al., 2008). Finally, the Chinese

4 Notably, the list of learner corpora on the Louvain site also includes the BAWE corpus of (proficient) undergraduate student writing from both L1 and L2 English writers.

Learner English Corpus (CLEC) contains 1 million words of English 'compositions' from senior secondary school students, English major and non-English major students in the PRC (Gui and Yang, 2003). In addition to corpora of Chinese students only, several corpora contain writing from students with different L1s, including Chinese. For example, the Corpus of English Essays Written by Asian University Students (CEEAUS) contains writing from students in the PRC, Japan, Korea and Taiwan. Essay choice is limited to two topics: either 'It is important for college students to have a part time job' or 'Smoking should be completely banned at all the restaurants in the country' (Ishikawa, 2010). All written texts in these corpora are extremely short; for example, students writing the in-class essays for WECCL are allowed just 40 minutes and are required to write at least 300 words. In CEEAUS, 20-40 minutes is allowed for the production of a minimum of 200 tokens and a maximum of 400 tokens. Frequently, however, publications from these corpora are available only in Chinese language publications (e.g. the Foreign Language Teaching and Research Journal) so my reading of these is limited to the abstracts (provided in English) and to summaries from Chinese colleagues.

The largest and best-known learner corpus is the International Corpus of Learner English (ICLE) (Granger et al., 2009) and this has produced a great many studies of general NNS writing (e.g. Gilquin and Paquot, 2008; Granger, 1998; Paquot, 2010; Petch-Tyson, 1998). Version two of ICLE contains over 3 million words of Part-Of-Speech (POS) and error-tagged writing by 'advanced' students of English as a Foreign Language (EFL) from 16 different first languages, including 'Chinese' (it is not clear whether this refers to Mandarin only or to a wider group of language users). 'Advanced' students are defined by the ICLE team as those in their third or fourth year of university undergraduate study in English, although this loose definition of proficiency has been critiqued as unhelpful by some ICLE users (e.g. Pendar and Chappelle, 2008; Chen, 2009). Texts are between 500 and 1,000 words long and are short, argumentative essays on a restricted range of general knowledge topics with titles such as 'Crime does not pay' and 'The role of censorship in Western society' (see discussion of titles in 2.6.2 and Appendix A for the complete list of suggested ICLE titles). Each subcorpus of essays is usually compiled by one or more academics at a single HEI per

country, meaning that an L1 subcorpus could consist solely of texts from a single cohort in one institution. Frequently, essays are effectively commissioned for the corpus by researchers asking student cohorts to write on assigned topics for a given time or within a given word range. The ICLE corpus is highly influential within NNS writing research and it is likely that its design has influenced later compilations of learner corpora (e.g. the Chinese corpora outlined previously).

In addition to corpora compiled by academic researchers, commercial corpora from publishers exist which include texts from varied groups of L2 writers (e.g. the Longman Learners' Corpus, Cambridge Learner Corpus). Test data from Cambridge ESOL are available to researchers on request and a small number of studies featuring Chinese student subcorpora exist (e.g. studies of Chinese and Greek students' IELTS papers by Mayor, 2006; and Mayor et al., 2007; and studies of Chinese and Spanish IELTS papers by Banerjee et al., 2007). Studies featuring mixed groups of NNS writing in test situations may also provide insights (e.g. Grant and Ginther's, 2000, research on the Test of Written English).

Reference corpora

Most researchers working with learner corpora use a comparative or reference corpus in order to contrast the writing of a particular group with that of another group. This comparison corpus is frequently one of expert academic writing such as the British National Corpus (BNC) (Burnard, 2007). The alternative choice for a reference corpus is to use a NS student writing corpus; the most widely-used of these is the Louvain Corpus of Native English Essays (LOCNESS[5]) (e.g. Guo, 2006, compares part of the Chinese learner corpus CLEC with LOCNESS). LOCNESS contains a total of 324,304 words of L1 English students' essays in three categories: British students' A level essays (60,209 words), British university undergraduate students' essays (95,695 words), and American university undergraduate

5 For further information on LOCNESS and other corpora developed at the Université Catholique de Louvain see: http://www.uclouvain.be/en-cecl.html.

students' essays (168,400 words). As with ICLE, the LOCNESS texts differ in terms of whether reference materials are used, whether they are timed or untimed, and whether they are argumentative essays on a general knowledge topic or literature essays (use of this reference corpus is critiqued in 2.6.2).

Summary

This section has given an overview of the literature available on Chinese students' academic writing. Although more studies of undergraduate writing are beginning to emerge from the BAWE corpus, most studies of Chinese students' writing (and on NNS writing more widely) are based on learner corpus data. While learner corpus texts are very different to assessed university writing, the findings may give insights into NNS academic writing in general and are discussed below.

Review question 2.b. What are the major findings from the research literature on Chinese students' writing?

To date, there are few studies of Chinese students' authentic undergraduate assignments in the UK. While this study concerns Chinese students' texts only, empirical studies using broader collections of learner corpus texts with varied L1s are also discussed in terms of how the findings might apply to Chinese undergraduate data. In this section a broad range of findings from the literature are first briefly presented; I then consider three selected areas in more detail which appear consistently across a wide range of the literature: informal items, preferred connectors, and the use of first person pronouns.

Overall findings

Many learner corpora studies have concentrated on grammatical 'errors' in L2 student writing, defining these 'problematic' areas against a NS norm (e.g. Chuang and Nesi, 2006; Wiktorsson, 2003). For example, Chuang and Nesi (2006) compiled a corpus of Chinese students' short, Business assignments on a foundation programme in the UK and tagged this for 'errors'; they estimated that over 10% of all identified 'errors' were 'missing' definite articles and 8.5% of 'errors' were 'redundant' definite articles. Surface-level features such as 'errors' of form are also discussed by Wiktorsson (2003: 67) (e.g. *believe on, at the other*

hand). A further broad area of discussion in the corpus literature on NNS writing is the 'overuse' of particular high frequency lexical items and chunks (e.g. Cobb, 2003; De Cock and Granger, 2004; Granger, 1998; Hinkel, 2003; Lee and Chen, 2009; Ringbom, 1998). For example, Lee and Chen (2009) found that texts from their study of Chinese undergraduate students in the PRC made greater use of a narrower range of lexical items and chunks than the NS students in the reference corpora (NS undergraduates in the UK and expert writers). Particular items discussed in their study are *according to, besides, we, the author, it can be seen* (some of which are further discussed in the sections below). Other studies have referred to the 'overuse' of vague, general nouns such as *people, things, man, woman, world, new, important* arguing that NNS students may not have the lexis required for greater specificity (e.g. Cobb, 2003; Granger, 1998; Hinkel, 2003).

A plausible reason for the extensive use of particular chunks by individual L2 writers is that these initially function as 'lexical teddy bears' (Hasselgren, 1994: 237), that is, frequently-used linguistic items which feel familiar and 'safe' (cf. Dechert's notion of 'islands of reliability' or 'fixed anchorage points', 1984: 227, in Granger, 1998: 156). As students widen their linguistic repertoire and become more confident, it might be expected that they would broaden their range of phrases. 'Overused' lexical items and chunks often include 'informal' or 'speech-like' items (e.g. *I think, to my mind*), connectors (e.g. *besides, first of all,* some of which may also be classified as 'speech-like'), and first person pronouns (i.e. *we, I*); these are each discussed below.

'Informal' items

'Informal' items refer to lexical words or chunks which appear to be informal in the context of academic writing (determined in this study according to descriptions in Biber et al., 1999). This phenomenon of informal written language is usually referred to in the literature as 'speech-like' items or as language with an 'oral tone' (e.g. Cobb, 2003; Field and Yip, 1992; Gilquin and Paquot, 2007; Granger, 1998; Hinkel, 2002, 2003; Lee and Chen, 2009; Mayor, 2006; Paquot, 2010) and is defined in the literature in comparison with a NS norm (such as the British National Corpus, academic writing section) or according to the researchers'

intuitions. In their comparison of ICLE and their reference corpus LOCNESS, Gilquin and Paquot (2007: 3) discuss 'spoken-like overused lexical items' such as *this is why, to my mind, really, by the way,* and *I want to talk about* which appear in the ICLE writing and seldom, if at all, in LOCNESS. The chunk *I think* is particularly singled out as an example of an expression which makes the L2 writers highly visible in their writing, and which is used far more frequently in speech than in academic writing. A similar set of 'speech-like' items given in Paquot's (2010: 151) related study are *that/this is why, look like, to my mind, from my point of view, of course, by the way, all in all.* An informal feel to academic writing may extend to a more dialogic style. In a study of 186 Chinese students' scripts from the International English Language Testing System proficiency tests (IELTS), Mayor (2006) found that these students had higher use of interrogative and imperative clauses than a comparable group of non-Chinese writers. This language use may in part be due to the nature of the task, since students are asked to give their opinion on topics of general knowledge. are frequently given in Chinese writing.

A reason commonly given for L2 writers' greater use of informal items is a reliance on more familiar, spoken language in their writing (e.g. Gilquin and Paquot, 2007; Hinkel, 2005; Paquot, 2010; Wiktorsson, 2003), hence the widespread use of the label 'speech-like'. Another possible explanation for the informal tendency, however, is the genre of the texts: most of the texts examined are short, argumentative essays with no research or preparatory reading. Regular reference to research articles or textbooks while writing reinforces academic writing style, as well as broadening the range of ideas beyond the student's own opinions. Students producing texts for the ICLE learner corpus may be asked to write about topics they have not previously considered from an academic standpoint, such as how much people should earn or the drawbacks of feminism, and may hold undeveloped views on these issues. It may be the case that ICLE contributors would write differently in academic essays on an academic topic area they are familiar with.

However, while a tendency to use more spoken-like language in writing may apply to ICLE studies featuring European students who are taught through a communicative approach and

who may have had exposure to spoken English, it seems less adequate as an explanation for Chinese students' usage. It is surprising, then, that Chinese writers are found to commonly use forms which are more prevalent in speech, given the lack of speaking practice in English language classes (as discussed in 1.2.3). Rather than being 'speech-like', it may instead be the case that informal lexical items and chunks are used across both spoken and written registers and that the issue is one of register discrimination.

Connectors

The second main area of agreement from the literature is the use of favoured connectors by NNS writers (e.g. Bolton et al., 2002; Hyland, 2008a; Lee and Chen, 2009; Milton, 1999). For Biber et al., (1999: 875), the main function of what they term 'linking adverbials' is 'to state the speaker/writer's perception of the relationship between two units of discourse.' Milton (1999: 226), using his own corpus of Hong Kong Chinese university student writing in English, explored the use of connectors in the writing and found that the students tended to 'overuse' the following chunks*: first of all, on the other hand, (as) we/you know, in my opinion*, *all in all,* particularly in sentence-initial position. In contrast, the L2 writers made little use of chunks which were used by L1 writers such as: *it can be seen that, an example of this is, this is not to say that*. The connectors used by the Hong Kong Chinese students appear more informal than those used by the L1 writers. Similarly, Lee and Chen (2009) comment on the use of the informal connectors *besides* and *what's more/what is more* in their corpus of Chinese undergraduate texts (from Linguistics students in a PRC university, named the Chinese Academic Written English corpus [CAWE]). Lee and Chen suggest that *besides* has an 'afterthought connotation' (p.288) and precedes less crucial information, arguing that it is not suitable for use at the start of a sentence or even paragraph in the way the Chinese students employ it (e.g. '…students' confidence might be increased. Besides, their interests might be stimulated…'). Lee and Chen also found that the somewhat informal expression *what's more/what is more* was used significantly more frequently by the Chinese writers than in the reference corpora of NS student writers and professional writers. Notably, however, as Lee and Chen's Chinese corpus consists of texts written in a PRC university, the different context may have contributed to the high use of particular linguistic items. It is likely that the

students' work would be graded by Chinese lecturers in Linguistics, for whom the 'overuse' (for Lee and Chen, when compared to NS or expert writers) might be less marked. Moreover, Lee and Chen compared the Chinese students' writing with Linguistics writing from both undergraduate and masters students in BAWE, yet the texts in BAWE are significantly shorter than the dissertations contained in CAWE (mean assignment length of 2330 words in BAWE and 5230 in CAWE). It may be the case that longer texts require proportionally more signposting in the form of connecting words and phrases, in order to guide the reader through the discourse.

Researchers disagree as to the effect of high use of specific connectors. Altenberg and Tapper (1998: 80) comment that '[r]elations that can be inferred from the text do not have to be marked explicitly, which means that a high frequency of connectors in a text does not necessarily improve its cohesive quality'. However, Galloway (2005: 338) makes the apt point that '[t]he use of discourse markers, excessive as it may be when statistically compared to native speakers' text, can contribute to the readability of less fluent writing', meaning that the use of more connectors may be beneficial. High use of connectors may also be 'closely connected with the individual writer's style and compositional technique' (Altenberg and Tapper, 1998: 83), that is, particular items may be preferred by individual students. Research into NNS writing has all too often assumed that higher, or lower, use of a particular linguistic feature is a 'problem' to be remedied, rather than simply a *different* way of meeting the challenge of academic writing. NNS students may favour particular connectors due to their familiarity (cf. Hasselgren, 1994); another possible reason offered by Milton (1999) concerns the tendency for ELT textbooks to provide lists of connectors, without distinguishing their use in different genres (this is explored further through an analysis of textbooks in 6.5).

Use of *we/I*

The third commonly-discussed characteristic of NNS writing considered in this section is the high use of first (and for some studies, second) person pronouns among L2 writers (e.g. Cobb, 2003; Lee and Chen, 2009; Lu, 2002; McCrostie, 2008; Petch-Tyson, 1998). For example, Petch-Tyson (1998) found that the NNSs in her study (French, Dutch, Swedish and

Finnish L1s) used first and second person pronouns between two and four times more frequently than the reference group of NSs (cf. findings in Cobb, 2003, and McCrostie, 2008). A keyword in Lee and Chen's (2009) study of Chinese undergraduate writers was the plural first person pronoun; this was often used within the lexical chunk *we can see* and functioned to direct the reader to a table or figure, or to organize the discussion (e.g. 'from the data in table 4, *we can see* that…').

However, this finding of greater use of pronouns is not confirmed in Hyland's (2002b) study of the use of personal pronouns in 64 Hong Kong student undergraduate theses. He compared the theses with a corpus of research articles, and found that the professional writers were four times more likely to use first person pronouns than the student writers. Instead of personal pronouns, students used passivization and *it*-clefts. Where students did self-refer, this tended to be in low-risk roles such as signposting and reflective sections discussing what they had personally gained from their project. Hyland (p.1110) suggests that presenting a strong authorial self 'clearly implies that the writer is a distinctive, individual creator… but this kind of identity is not shared by all cultures'. He argues that the student writers' unwillingness to commit to their views 'may, in part, be a product of a culturally and socially constructed view of self which makes assertion difficult' (p.1111) (though note that Hyland's later work found differentiation between Hong Kong writers within different disciplines e.g. 2008a,b).

Summary

Research into Chinese students' academic writing, and into NNS writing in general, has highlighted language items which, in Granger's (2004: 132) words, are considered to be 'either over- or underused by learners and therefore contribute to the foreign-soundingness of advanced interlanguage'. The literature survey reveals general agreement that these writers make high use of particular lexical items and chunks; these may include 'informal' or 'speech-like' items and particular connectors. Findings on first person pronouns are less clear with most studies showing a high use of *we* and *I,* though Hyland's (2002) study on Hong Kong students indicates low use. However, much of the research on Chinese and other L2 students' academic writing is conducted on short, argumentative essays and

findings may not be applicable beyond these. Moreover, the reference corpora used are not always directly comparable, and instances of 'overuse' may be due in part to differences in genre rather than the impact of the L1 and the L1 environment.

Review question 2.c. To what extent has this research area been covered?

From this review of literature on Chinese students' writing, it seems that although there has been a certain amount of work done on Chinese students' writing and NNS writing more generally, this has focused on short argumentative assignments within learner corpora, or on postgraduate writing, and there has been little research to date on Chinese students' undergraduate writing. While findings from the many multiple L1 learner corpus studies offer insights into NNS writing in English in general, a major drawback to learner corpus studies is the restricted genre of the writing. Since the short, argumentative essay typical of learner corpus texts is not equivalent to the varied, discipline-specific writing in undergraduate degrees, it is unclear how far the features of NNS writing identified in these studies are relevant (this point is discussed more fully in 2.6).

In the study reported here, Chinese and English student writing is confined to UK undergraduate writing and includes texts from a variety of genres (e.g. case studies, reports, reflective writing, laboratory reports). Unlike Chen's (2009) study (which also uses BAWE data), only undergraduate texts are selected, enabling stronger claims to be made as to the nature of degree-level writing in the UK.

2.4 Variation over time

Review Question 3: According to the research literature, how do the identified characteristics of Chinese students' undergraduate writing in UK universities vary over the period of undergraduate study?

Review question 3.a. What is known about variation in academic writing by English students at different levels of undergraduate study?

The question of variation in English students' writing across undergraduate year groups is asked here in order to set findings on the variation of Chinese students in context. Of the (relatively low) number of studies of undergraduate writing by L1 English students, few have taken a longitudinal approach and compared writing over time; however, some studies have compared data collected at the same point in time from different students at varying stages of degree study (e.g. Hewings, 1999, employing her own data; Gardner, 2008, and Thompson, 2009, both using BAWE corpus data). Granger (2008: 11) terms these studies 'quasi-longitudinal' since the same students are not followed, but the underlying assumption is that students' writing changes over time. The studies by Hewings, Gardner, and Thompson are considered below.

Hewings (1999) compares writing by Geography undergraduates in year 1 and year 3, focusing her analysis predominantly on the beginnings of independent clauses. She found that while first year students emphasize real world phenomena as they demonstrate their knowledge of course content, year 3 students 'focus on disciplinary knowledge-making practices in the form of reporting and evaluating the work of others' (p.212). The increased engagement with the arguments of the discipline of year 3 Geography students became apparent through their 'greater awareness of disciplinary conventions regarding use of maps and diagrams as evidence rather than as illustrations of processes' (p.212). Hewings' focus on anticipatory *it* clauses (e.g. *it is estimated that, it is possible that*) revealed that third year students made greater use of these structures, as a way of indicating their own opinions in a less overt way than employing personal pronouns. She suggests that the higher use of *it*-clauses and the higher commitment shown through modality 'may be interpreted as indicating a greater concern with argument and evaluation' (p.247). Hewings' study suggests there is a clear variation in writing from a more categorical style in year 1 to a more nuanced and sophisticated one in year 3.

Gardner (2008) also found variation across undergraduate levels, with year 1 History students focusing on the *phenomenon* of 'History' in year 1 and year 3 students discussing the *metaphenomenon* of being a 'historian', the variation suggesting a move from more factual knowledge to discussion of this knowledge. In her study of History and Engineering assignments in the BAWE corpus, Gardner comments on what she terms 'development' (p.17) in text length from year 1 to year 3 (measured through the mean assignment length in words) with assignments in each discipline steadily increasing across the three year groups. In Engineering, Gardner notes a progression in writing within genres such that assignments require specific types of writing which are then combined, 'culminating in long projects and design proposals where multiple skills are integrated' (pp28-9). The variation in History genres is much slighter, with the essay dominating throughout, leading Gardner to comment that '[t]he demands on engineers to develop a range of writing skills are therefore much greater than on historians who have more opportunity to hone specific argumentation skills in a narrower range of genres' (p.29). While essays in History are no doubt challenging, and tend to be *longer* than Engineering assignments, the undergraduate History student can learn from essay feedback about how to write in this particular genre whereas the Engineering student must meet the challenge of writing in multiple genres such as reports, reflective pieces, proposals and methodology recounts. Despite this challenge, Engineering is one of the most popular choices of discipline for Chinese students in UK universities.

Thompson (2009) also focuses on the disciplines of History and Engineering in BAWE and explores first person pronouns, four-word chunks, and the pattern *it* + be + ADJ + *to/that* across disciplines and year groups. Regarding use of the first person singular pronoun he comments that for both disciplines the use of *I* rises slightly from year 1 to year 2 before falling overall by year 3. In History, Thompson analyzes year 1 and year 3 assignments from the same student, and finds evidence to suggest a move from use of *I* to convey what he terms 'local decisions' (p.63) (e.g. 'I have chosen 1757 as the beginning') to use of *I* to structure the argument (e.g. 'I would even argue that'). Thompson's extraction of four-word lexical chunks reveals variation across year groups in Engineering with a move from chunks related to laboratory experiments and to Engineering concepts in year 1 (e.g. *the moment of*

inertia), through to more referential and procedural chunks in year 2 (e.g. *it can be seen to, can be used to*) which continue into year 3 and are joined by chunks expressing causality (e.g. *as a result of*). In contrast, he found little evidence of a progression of chunks in History, with variation due instead to localized factors such as modules on particular time periods. Finally, the study found an increase in use of the pattern *it* + be + ADJ + *to/that* in year 3 for both disciplines, as the student writers comment more on propositions or processes. Thompson sees this increasing use of this particular pattern as 'an indication of a growing ability to express judgements within one's writing in an authoritative manner' (p.79). For both History and Engineering, then, evidence was found for variation across the year groups, though the nature of this variation differs according to the discipline.

This section has reported on findings from three studies on NS undergraduate students' writing in terms of the variation shown in assignments produced from year 1 to year 3, though it should be borne in mind that none of the studies are 'true' longitudinal ones since the students in each year group are (mainly) different individuals. A theme of all the studies is the change in students' writing with a gradual progression as they acquire the expected writing conventions of particular disciplines. This progression includes a move from use of simple topical Themes (i.e. relating to the subject of the sentence) in year 1 to multiple Themes (i.e. a combination of interpersonal, textual and topical Themes at the sentence start) in year 3 (Geography); from a display of factual knowledge to an increase in discussion of the arguments surrounding this knowledge (Geography, History); a reduction in the use of *I* by year 3 (History, Engineering); a change in the types of lexical chunks employed (Engineering) and from shorter to longer assignments (History, Geography, Engineering) in an increasing range of genres (Engineering). It seems from all three of these studies that variation across undergraduate levels is inextricably linked with what it means to learn to write within a particular discipline; this is further explored below.

Review question 3.b. What literature is available which traces variation in academic writing by Chinese students at different levels of undergraduate study?

The majority of learner corpora are synchronic, that is, they are collections of texts written at a particular time and as such represent a snapshot of learner writing (e.g. ICLE). Texts for

this type of corpus are much easier to collect than for a diachronic, or longitudinal, corpus tracing development over time from the same students. The longitudinal studies which exist frequently follow the writing of a small number of students, adopting a case study approach (e.g. Li and Schmitt, 2009, trace the writing development of one Chinese masters student in a UK university, and Larsen-Freeman, 2006, examines the variation in five Chinese students' written narratives over a six-month period in the US). Kinzley (2011) studied 99 texts from eleven '2 + 2' students from partner institutions in the PRC undertaking years 2 and 3 of a Media and Cultural Studies degree course at Lancaster University. He traced the students' learning trajectory through a pre-sessional course and to the end of term 2 of their first year at Lancaster with respect to student uptake of the rhetorical devices which were taught.

As in Hewings' study of L1 English students of Geography, Wiktorsson (2003) takes a quasi-longitudinal approach with her corpora of Swedish students' writing in English. She compares the writing of Swedish Upper Secondary Students (the SUSS corpus), and the Swedish subcomponent of ICLE (SWICLE). Although not concerning Chinese students, Wiktorsson's study is briefly discussed within this review question as this is a comprehensive study of variation across year groups. Similarly, though at postgraduate level, Hyland (2008b) compares cohorts of masters and PhD level writers in Hong Kong within a range of disciplines, assuming a trajectory of academic writing from these students to professional academic writing. Other studies have compared less proficient texts and more proficient ones, with the assumption that there is a linear development over time such that students become more linguistically proficient (e.g. Kennedy and Thorp's, 2007, study of low and high-scoring IELTS papers; and Grant and Ginther's, 2000, study of low and high-scoring Test of Written English papers).

Given the dearth of longitudinal research in the true sense of the term, the next section considers both studies of development over time, whether longitudinal or quasi-longitudinal (e.g. Wiktorsson, 2003), and also studies of low and high proficiency (which may equate to development over a period of time) (e.g. Grant and Ginther, 2000).

Review question 3.c. What are the major findings from the research literature on variation in academic writing by Chinese students at different levels of undergraduate study?

It is often assumed that language development follows a path from simple to more complex language, and from less varied to more varied lexis; these features have been measured in studies of NNS writing through aspects such as the quantity of language and the variety of lexical chunks produced. The studies below employ a range of corpus linguistic procedures to illustrate variation across levels of academic writing.

Increase in fluency, accuracy and complexity

Larsen-Freeman's (2006) study of five students from the PRC found that over a six-month period, participants' writing became more fluent, accurate, and complex, as measured through quantitative means such as the number of words, 'errors' and clauses per 't-unit' (an independent clause, possibly containing additional dependent clauses, phrases or words) and through qualitative analysis of the 'idea units' of the narratives. However, the students were required to both write and orally recount exactly the same narrative every six weeks, meaning that linguistic 'development' may be at least in part due to the effects of task repetition. Larsen-Freeman recognizes this inherent problem, citing Bygate et al.'s, (2001) discussion of how complexity and fluency, though not accuracy, are improved when a task is repeated. Nevertheless, she argues that improvement in a task means that the student 'has developed greater language resources with which to accomplish the task' (p.613).

Also measuring fluency and variety, Grant and Ginther (2000) compared a range of NNS writers' test papers at varying levels of proficiency (the L1s are not specified) in terms of the mean number of words per essay, the range of vocabulary used (measured through the type/token ratio or number of distinct words divided by the number of running words in the corpus), and the average word length (measured in alphanumeric characters). They found that for all three measures, the more proficient writers (as measured through their Test of Written English scores) compose 'longer timed essays with more unique (and on average, longer) word choices' (p.140). Grant and Ginther also calculated mean rates for a variety of linguistic features in the essays (e.g. first person pronouns, modals of necessity) but,

unfortunately, they give their figures per essay rather than through a normalized wordcount, meaning that these cannot be easily compared across essays. They then comment that the use of various parts of speech increases as proficiency rises, adding that '[t]his is not surprising because, as we have already noted above, the writers are producing more language' (p.135).

Variation in lexical chunks

Many researchers have measured development in terms of the lexical chunks used by writers (e.g. Larsen-Freeman, 2006; Paquot, 2010; Wiktorsson, 2003), with progression equating to a move from a smaller to a wider range of chunks (whether classified structurally or functionally) or from more speech-like to more academic, information-dense chunks. In her comprehensive study of Swedish students at different levels of academic study, Wiktorsson (2003) found that the lower level secondary school students used fewer chunks than either the 'advanced' university students or the NS groups. However, while the 'advanced' students used equal numbers of chunks to native speakers, they made greater use of chunks more commonly found in speech than in writing. Wiktorsson also identifies instances of what she terms 'erroneous target language prefabs' (p.158) where students aim to produce an idiomatic chunk but create an unidiomatic one (e.g. *one in a while, at the other hand)*. These erroneous chunks are produced mainly by the secondary students, suggesting that as language develops, students learn more 'nativelike' expression. It may be the case that these non-nativelike chunks are learned initially, and then become a chunk for the NNS (cf. later discussion of Hoey's, 2005, theory of lexical priming in 3.2). Although Wiktorsson's study concerns only Swedish students, her study revealed interesting findings which may be pertinent to other L1 groups.

Hyland's (2008a) study of masters theses, PhD dissertations and professional academic research articles also concerns lexical chunks (which he calls 'clusters') and he argues that these 'come to signal competent participation in a given community of users' such that 'an absence of such clusters might reveal the lack of fluency of a novice or newcomer to that community' (p.2). In contrast to other studies of NNS writers, Hyland confines his discussion to disciplinarity and level of writer (master, PhD or professional academic) and only briefly

discusses the fact that his students are Hong Kong Chinese. For instance he states that the masters students' low use of clusters which focus on the writer or reader of the text (e.g. *are likely to be, as can be seen,* termed 'participant-oriented') 'mirrors Hong Kong students' preference for author anonymity' (p.54). Hyland links this to his 2002 study which highlighted Hong Kong students' low use of first person pronouns. Hyland (2008b) found evidence of a cline of cluster usage from masters' theses to PhD dissertations to research articles, such that clusters concerned with references to the research ('research-oriented') decreased in the longer genres, whereas clusters concerned with the organization of the text ('text-oriented') or with the reader or writer's views ('participant-oriented') increased. However, a limitation of Hyland's study is his categorization of clusters into a single functional group. He assumes that a cluster has one primary function, no matter which group of writers uses it, and does not discuss any encountered difficulties in either assigning these categories or in this assumed monofunctionality (this point is discussed further in 3.5.2). Despite the difficulties, the categorization of clusters or chunks according to the function they carry out provides a useful means of comparing texts from different student groups (cf. studies employing functional categorization by Cortes, 2004; Scott and Tribble, 2006).

This section has indicated that NNS' writing develops over time in that texts become longer, more accurate, and employ more varied lexis, with the development of lexical chunks being a particular focus of several studies. However, none of the studies have measured variation of a substantive group of Chinese students' texts.

Review question 3.d. To what extent has this research area been adequately covered?

As pointed out in the discussion of 3.a above, large-scale, diachronic studies of student progression are seldom carried out, due to the difficulties in collecting longitudinal data from the same group of learners over time. Kinzley (2011: 376) argues that such 'tracking studies' are necessary in order to improve pre-sessional courses, suggesting that these could follow the academic progress of students from the beginning of the pre-sessional to at least the end of their second term of undergraduate study. He distinguishes between the *linguistic* proficiency emphasized on many preparatory courses and the *discipline-specific* proficiency

required within undergraduate study. While my study is 'quasi-longitudinal' rather than being a 'true' longitudinal study, it will aim to offer insights into variation across the year groups.

To counter this lack of longitudinal studies, a longitudinal database of L2 English writing is under construction at the Université Catholique de Louvain and will comprise writing from students across a range of L1s (LONGDALE, Louvain Corpus of Native English Essays[6]). Although findings from learner corpora cannot be automatically extended beyond the genre of the short essay, the development of this new corpus may offer useful insights into longitudinal variation in NNS writing.

One area in which student writing may show progress is in the development of disciplinary-specific writing. It is likely that year 3 students, having spent three years studying, are closer to writing in ways which are valued within their discipline (Hewings, 1999), and the next section explores the literature within this area.

2.5 Disciplinarity

Review Question 4: According to the research literature, how are the identified characteristics of Chinese students' undergraduate writing in UK universities affected by writing conventions in different disciplines?

Review question 4.a. How can the concept of a 'discipline' best be defined and operationalized for the purposes of this study?

Hewings (1999: 35) asserts that 'what disciplines write about and how they write about it is central to their construction and maintenance of identity' and similarly, Hyland and Tse (2007: 240) argue that 'all disciplines shape words for their own uses'. In writing as members of an academic community, students are thus 'learning new discourse practices and assuming different conceptions of themselves as writers' (p.37). Before considering

6 For further information on LONGDALE and other corpora developed at the Université Catholique de Louvain see: http://www.uclouvain.be/en-cecl.html.

disciplinary discourses, however, it is first necessary to define what is meant by the concept of an academic 'discipline' (cf. Nesi et al., 2005).

An established means of assigning common features to disciplines, and thereby defining and categorizing them, is through the dimensions of 'hard-soft' and 'pure-applied' first discussed by Biglan (1973) and Kolb (1981) and followed up by Becher (1989, 2004) (see Figure 2.1).

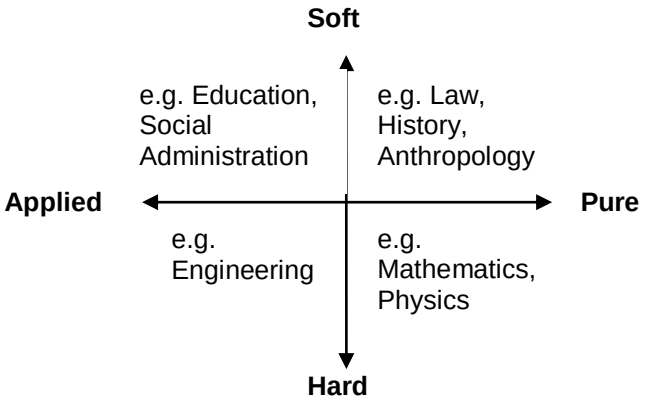


Figure 2.1 Soft-hard, pure-applied discipline paradigm (after Becher, 1989, 2004)

Thus, each discipline, or group of disciplines can be placed along the 'soft-hard' and the 'pure-applied' dimensions, relative to other disciplines. Drawing on the earlier work of Biglan (1973) and Kolb (1981), Becher (1989) suggests that 'hard-pure' disciplines include Mathematics and the Natural Sciences with their well-defined boundaries of knowledge and the cumulative growth of findings. In contrast, the 'soft-pure' disciplines of Law and other Humanities have more permeable boundaries between academic 'territories' (Becher, 1989) and are characterized by contention over proposed theories. In the applied dimension, 'hard-applied' disciplines are concerned with control over the physical world, as typified by Engineering. 'Soft-applied' disciplines are based on interpreted knowledge and how this relates to human society, as exemplified by the disciplines of Education and Social Administration.

However, there is far from uniform agreement as to where to place individual disciplines within a 'hard-pure' and 'soft-applied' matrix. For example Neumann et al. (2002: 407) suggest that a discipline may 'straddle two categories' (e.g. Biology contains both 'soft pure'

and ‘hard pure’ elements), may contain ‘deviant specialisms’ (e.g. the area of sociometrics within the discipline of Sociology), or may change over time. Moreover, there is often a lack of agreement among discipline specialists themselves as to where a discipline belongs along the ‘hard-soft’ and ‘pure-applied’ dimensions. For instance the relatively new discipline of Computer Science is generally held to be a ‘hard-applied’ discipline (alongside Engineering), yet has at times been classed as ‘soft-pure’ (alongside pure Mathematics) (Clark, 2003). Despite difficulties in classification, however, the ‘hard-soft’, ‘pure-applied’ matrix has a usefulness in placing disciplines in groups and thereby considering similarities and differences within, for example, the requirements for assessed student writing.

In the BAWE corpus a ‘discipline’ is framed in terms of a subject taught within one department. This is a practical solution to the problem of boundary-setting, and works in part due to the initial BAWE practice of assigning data collection for each discipline to a single university. However, as a general principle this definition of discipline is both contentious, since each institution may set the boundaries slightly differently, and is in a state of flux, since departmental knowledge areas are continually expanding and ‘new’ disciplines may be formed from within existing ones. While recognizing these tensions in determining discipline boundaries, as this study primarily uses BAWE corpus data I adopt this same alignment of discipline with the departments of participating universities.

Review question 4.b. What is known about how writing varies within different academic disciplines by English students at undergraduate level?

As previously discussed, few corpus studies prior to the release of the BAWE corpus have focused on large-scale analysis of undergraduate writing across disciplines. A range of available studies of English students’ undergraduate texts were outlined in response to review question 1.a in 2.2. (e.g. Bruce, 2010; Gardner, 2008; Gardner and Holmes, 2009; Jung, 2011; McKenny, 2005; Nathan, 2010; Nesi and Gardner, forthcoming, 2011; Thompson, 2009). This section draws on studies by Hewings and North (2006), Bruce (2010), Gardner (2008), and Thompson (2009) in order to bring out salient points about the effect of disciplinarity on student writing.

Hewings and North (2006) compared corpora of undergraduate student writing in Geography and History of Science, using Halliday's (1994) thematic analysis. The Theme or 'point of departure for the message' (Halliday, 1994: 34) for each sentence was classified as 'simple' (subject Theme only, e.g. *the tsetse fly*), 'multiple' ('textual' and/or 'interpersonal', e.g. *Firstly, we*…) or 'marked' (e.g. containing an initial circumstantial adjunct such as *in the Cumberland Basin, Australia*). The results showing a clear divide between the two disciplines with the History of Science texts containing a higher percentage of marked Themes than the Geography texts. Moreover, a difference was found between History of Science students who came from an 'arts' background and those from a 'science' background, with the former making greater use of both interpersonal and textual Themes. Hewings and North (p. 271) conclude firstly, that student writing 'demonstrates at least some disciplinary variation, such that what is valued in one context cannot be assumed to be valued equally in another context' and secondly, that students' writing in one discipline may be influenced by the writing they have carried out in a previous discipline. This is significant as writing within different disciplines is becoming an even greater issue in current modular degree courses where students undertake several areas of study. It may be difficult therefore for students to both fully understand what kind of writing is valued within a discipline, and to disentangle the effects of previously-studied disciplines.

Added to the challenge of writing within several disciplines, either concurrently or sequentially, is the issue of genres of student writing within these disciplines. The same nomenclature for genres may be used across disciplines, but may not in fact relate to the same conventions. Bruce (2010) compares essays within Sociology and English, using data from the BAWE corpus in a study of the rhetorical purposes of the genre. The investigation into the essay genre within each discipline uses Bruce's own genre analysis framework and draws on Hyland, 2005a, giving clear labelling of the rhetorical and textual resources used in each of the 20 essays. Bruce concludes that the essays had 'surface similarities' (p.162) such as including an introduction, body and conclusion, but that significant differences exist between the two sets of essays. The Sociology essays were found to have a more developed move structure at the start of each essay, with more metadiscoursal mapping

throughout. In contrast, the English essays employed less metadiscourse, leading Bruce to state that these assignments assumed greater reader responsibility. Bruce's study of essays within two disciplines leads him to conclude that 'despite common naming, assignment tasks across disciplines will vary considerably in terms of the type of writing required' (p.163). However, Bruce's essay sample is very small, using just ten essays from each discipline, and he does not attempt to balance the two (randomly-selected) datasets by year group (Sociology has just one text from a year 3 student whereas English has six), or by grade (three of the Sociology essays receive the 'distinction' grade but all ten English essays are in this classification). Moreover, the 20 essays considered represent work from just 11 individuals; given that he required only a subset of the available data, it would be reasonable to extract essays on the basis of one text per individual to reduce the effect of idiosyncrasies. This latter point perhaps illustrates the difficulty in working with an unfamiliar dataset: Bruce does not appear to be aware of the ID conventions for BAWE corpus texts wherein the same number is awarded to each individual with a different letter denoting each text (e.g. 3006a and 3006e are different texts by the same student).

Again using BAWE corpus data and focusing primarily on genre, Gardner (2008) studied writing in History and Engineering through a variety of means (ethnography, multidimensional analysis, Corpus Linguistics, Systemic Functional Linguistics). She comments on the range of genres that undergraduate students have to contend with (as discussed in 2.4). Thus, Engineering students have to produce reflective writing, writing for non-specialists and writing for professional purposes (cf. Nesi and Gardner, 2006; Leedham, 2009). Gardner comments that 'much assessed writing in Engineering aims to produce professional engineers' (p.27), hence genres in Engineering include design plans in response to a specific design brief, and design proposals which include costings of a project. Essays and reflective writing are also required and may discuss Engineering ethics or narrate the process of a group project. Gardner discusses how '[I]n stark contrast, writing in History focuses almost exclusively on the essay' (p.11); however, these essay assignments were found to be longer than the mean average assignment length in Engineering (mean word length of 2654 in History and 2310 in Engineering). Gardner also compares automated

counts of tagged features which reveal that the number of tables, figures, formulae and lists is far greater in Engineering than in History, though she does not discuss the impact of this writing on either the text lengths or genre classification. Disciplinary differences such as different genres and the use of visual features and list-writing might be expected in disciplines at opposite ends of Becher's (1989) dimensions (History is a 'soft-pure' discipline and Engineering is 'hard-applied'), but the conventions are seldom discussed in either EAP or subject-specific classes with both NS and NNS students expected to learn how to write within their discipline with minimal outside assistance.

In his study of History and Engineering from the BAWE corpus Thompson (2009) examined the use of first person pronouns and found that Engineering students employ both inclusive and exclusive *we* (e.g. 'We can see that FORMULA is proportional to…'; '…before we iron out all the problems…'). In comparison, History students use only the authorial, inclusive sense of *we* (e.g. 'Here we find the distinction that…'). The disciplines also differ in their use of *I*, with Engineering students projecting different identities through their writing as they convey a sense of themselves as an actor (e.g. *I will use*), a reflecter (e.g. *I have learnt*), and professional (e.g. *I propose*) (Thompson makes use of Tang and John's, 1999, six categories of first person pronouns; these are discussed further in 5.5). In History the main genre is the essay (as Gardner, 2008, also observes) with a consequent need for organizing prose (e.g. *I will explore*); the other use of *I* is in the student writer's projection of their identity as 'an historian and an arguer' (e.g. *I would* argue) (Thompson, 2009: 63). However, one aspect which Thompson only briefly alludes to is the higher proportion of NNS writers in the Engineering subcorpus of BAWE in comparison with History. In Engineering approximately 65% of texts are produced by L1 English students, with the remainder written by L1 Chinese, Arabic, Tamil, Sinhala, Gujarati and Hindi speakers; in History the proportion is 93% L1 English speakers, with the remainder Dutch, Hindi and German. This higher proportion of NNS writers in Engineering, while still resulting in proficient writing, may also contribute towards differences in the use of *we/I* across the two disciplines. Engineering is one of the most popular disciplines for Chinese students in BAWE, and texts from this discipline are explored in Chapter 7.

This section has discussed a sample of the available literature on L1 English students' undergraduate texts, finding that the effect of both discipline and genre have a great impact on the writing. The following sections turn to the smaller set of data concerning characteristic of Chinese students' writing across disciplines.

Review question 4.c. What literature is available on Chinese students' writing in different disciplines?

Excepting studies of postgraduate writing (e.g. Hyland, 2008a,b), research into Chinese students' writing has been limited to either a single discipline (e.g. Flowerdew, 2003; Kinzley, 2011; Lee and Chen, 2009; Nathan, 2010) or to learner corpus studies of general topics. However, studies have tended to make comparisons between the Chinese students' writing within a specific discipline and NS student and/or expert writing within the same discipline, rather than comparing writing in one discipline with that in another. For example, Lee and Chen (2009) compare keywords and key chunks in three corpora within the discipline of Linguistics: Chinese students' undergraduate dissertations from a PRC university (from CAWE), UK undergraduate assignments (from BAWE), and research articles. However, Lee and Chen barely comment on the effect of discipline. It seems that studies which limit the data from Chinese students to a single discipline have done so either for ease of collection or to control the number of variables. Consequently, findings tend to be given at a general level of academic writing and little comment is made as to aspects of disciplinarity. (It should be noted here that in the PRC, as in other L2 English environments, undergraduate students are not required to write in English within a discipline unless their major is English or Linguistics).

Future research: VESPA

The development of a new corpus by the creators of ICLE may give rise to more discipline-specific research within and across L2 groups. The aim behind the Varieties of English for Specific Purposes dAtabase (VESPA[7]) learner corpus is to collect examples of writing by

7 For further information on VESPA and other corpora developed at the Université Catholique de Louvain see: http://www.uclouvain.be/en-cecl.html.

different L1 groups, from as many disciplines as possible and ranging from first year undergraduates to PhD students. Currently (May 2011), VESPA has partnerships with institutions in five countries within Europe. As most of the collaborating departments are English language or Linguistics, it seems likely that these disciplines will be completed first. Given the complexity of negotiating access to other academic departments, and persuading students to donate their assignments (as there is no budget to pay contributors), the compilation of a comprehensive database is likely to take a considerable amount of time.

Review question 4.d. What does the literature tell us about Chinese students' writing in different disciplines?

Review question 4.c. pointed to the lack of discussion on disciplinarity in Chinese students' writing, commenting that studies have instead primarily compared NNS texts with NS texts. In their study of writing in Linguistics, Lee and Chen (2009) do not comment on disciplinary aspects of the writing, beyond a few remarks on topic bundles (e.g. the chunk *the teacher should* seems particular to the L2 students' topics), concentrating their discussion on the 'problematic' language of the Chinese students in CAWE, terming this corpus of undergraduate texts a 'learner corpus'. Given the lack of research on disciplinarity in Chinese students' writing, this section confines the discussion to a postgraduate study by Hyland (2008a).

As noted in 2.3, Hyland's twin concerns are with disciplinarity and level of writer (masters, PhD or professional academic) and he only briefly discusses the fact that the writers are Cantonese-speakers from Hong Kong universities. Hyland's (2008a) study of 3.5 million words from four disciplines (Applied Linguistics, Biology, Business Studies and Electrical Engineering) combines data from postgraduate (masters and PhD) L1 Cantonese students with that of professional research articles (L1s are not given for these but are presumably varied though articles are likely to have had considerable input from reviewers and editors). Hyland's focus is on four-word chunks (which he named 'bundles') which he extracts on the basis of frequency. He comments that bundles such as *as a result of* and *as can be seen* will

be familiar to ‘writers and readers who regularly participate in a particular discourse’ (in this case, academic), and remarks that ‘[c]onversely, the absence of such clusters might reveal the lack of fluency of a novice or newcomer to that community’ (p.5). As such, the use of appropriate clusters could demarcate conformation to disciplinary conventions. Hyland’s discussion contains some slippage in his references to ‘students’ and NSs, for example he asserts that ‘it is often a failure to use native-like formulaic sequences which identifies students as outsiders’ (p.7). Here, the reader is led to believe that the ‘students’ referred to are general NNSs (L1 Cantonese), whereas the broad class of ‘students’ include NSs *and* NNSs.

Hyland categorized the extracted four-word bundles as either ‘research-oriented’, ‘text-oriented’ or ‘participant-oriented’ (see discussion of these terms and the issue of a monofunctional classification in 3.5.2). He found distinct differences in the use of the bundles across discipline groups, with Biology and Electrical Engineering having a greater concentration of research-oriented bundles (e.g. *in the presence of*), and suggests that this illustrates the emphasis in science of ‘the empirical over the interpretive’ (p.15). Applied Linguistics and Business, in contrast, have a high number of text-oriented bundles (e.g. *in the sense that*), which in Hyland’s view ‘reflects the more discursive and evaluative patterns of argument in the soft knowledge fields’ (p.16). Participant-oriented bundles (e.g. *are more likely to*) were also more common in the ‘soft’ fields of Applied Linguistics and Business, and indicate writers’ wish to establish claims, evaluate and engage the reader. Here, Hyland points out that participant-oriented bundles were more common in the research articles, remarking that avoidance of these may ‘perhaps reflect the influence of a second language factor’ (p.19).

Hyland concludes from his study that ‘writers in different fields draw on different resources to develop their arguments, establish their credibility and persuade their readers’ (p.20). While his study does not primarily focus on the L1 of the majority of the texts (Cantonese), his discussion of four-word bundles has much to offer an investigation into disciplinary difference. Potential shortcomings in the categorizations used are discussed in 3.5.2.

Review question 4.e. To what extent has this research area been adequately covered?

Section 2.5 has outlined the available literature concerning English students and Chinese students' writing within different disciplines. The literature review has revealed that while it is relatively common for a study to limit a dataset to one discipline, it is not common to then compare this discipline with another and thus research the effect of disciplinarity on (in this case) Chinese students' writing.

2.6 Discussion and implications

Sections 2.2 – 2.5 have suggested that undergraduate assignments are a far from monolithic entity, echoing Moore and Morton's (2005: 54) comments on the 'great diversity' of this genre grouping. The reported studies on both English and Chinese students' assessed undergraduate writing have illustrated the complex and multi-genred nature of these texts, and the learning process undergone by all students, NS or NNS, in understanding disciplinary conventions. Given the scale of Chinese students' presence in UK universities, it might be expected that there would be a considerable body of research into this group's academic writing at all levels. However, the majority of research studies are limited to learner corpora; moreover postgraduate writing has been studied more than undergraduate. In addition, Hong Kong Chinese students have perhaps been over-represented in studies featuring 'Chinese' students (e.g. studies by Flowerdew, 2003; Hyland, 2002, 2008a,b; Milton, 1999) while the writing of students from the People's Republic of China is under-explored. Additionally, existing studies of Chinese students' writing overwhelmingly take a deficit approach and contrast the texts with a NS 'norm'. This section discusses this dominant deficit approach, contrasting this with the more descriptive, academic literacies approach taken by this study. The established ICLE corpus is compared with the more recently created BAWE corpus to illustrate the differences in compilation and what each can offer to research into student writing.

2.6.1 Deficit versus descriptive perspectives

The studies reviewed in 2.2 – 2.5 of Chinese student writing, and NNS writing more generally, adopt a *deficit* approach in that NNS writing is viewed as something to be 'improved' or 'corrected'. NNS writing is compared to either a NS student 'norm' or to professional academic writing and, in either case, is seen to fall short of these 'standards'. Often, the language used in learner corpora studies is couched in terms of a deficit discourse surrounding NNS writers rather than one of variational 'difference'. For example De Cock (2000: 65) comments on the 'foreign-soundingness' of NNSs' speech and writing with its 'overuse', 'underuse' and 'misuse' of chunks, its 'stylistic deficiency' (p.58)[8], and Gilquin and Paquot (2008: 58) conclude that 'remedial materials' are required to help NNSs 'overcome register-related problems'. Similarly, Chen and Baker (2010: 34) discuss 'immature student academic writing… [across] three groups of different writing *proficiency levels'* in their corpora of NNS student, NS student and expert academic writing (emphasis added). This dominant deficit model of student writing which prevails in most corpus studies of student writing implies that, at some point, both NS and NNS will have acquired the target discourse of academic writing, both generally and in a specific discipline. That is, there is an endpoint to the lengthy process of learning to write in an appropriate academic style and, once this is reached, writers are able to take a full part in academic dialogue. Reference is made to the existence of a linguistic proficiency cline from low-level NNSs to high-level NNSs through to NSs and culminating in the language of professional academic writers, at which point the NS/NNS divide ceases to be significant (e.g. Chen, 2009; Chen and Baker, 2010). Belief in this ability cline assumes that student writers, and particularly NNS student writers, encounter 'problems' in academic writing which need intervention to enable the student to correct them.

In contrast to the 'deficit perspective', the alternative academic literacies perspective views writing within the academy as a set of social practices in which genre, context and culture are highly significant, highlighting 'the variety and specificity of institutional practices, and students' struggles to make sense of these' (Lea and Street, 2006: 376). The institutional

8 The use of the terms 'over/under/mis use' are prevalent in the learner corpus literature, though note that Gilquin et al. (2007: 322) describes them as 'descriptive, not prescriptive terms'.

nature of what counts as knowledge within a disciplinary setting acts as a gatekeeper, restricting access to the academy. As this challenge is not unique to NNSs but is also applicable to NSs, an academic literacies perspective does not dichotomize NS and NNS students but instead views all university students as learning how to write within the academy. Academic literacies highlights the institutional nature of what counts as knowledge within both the academy and within a discipline. Writing in an acceptable way is viewed as a situated process which is constantly in flux, since the epistemology of disciplines is complex and dynamic. Only the academic literacies model focuses on specific contexts of student writing such as writing in response to a particular assignment prompt for an individual lecturer within a department in an institution.

For both NS and NNS writers, the academic literacies model regards writing in UK HE as problematic, and in a state of flux with students taking part in a constant struggle to establish the preferred ways of making meaning within their particular context (Lea, 2004; Lea and Street, 2006; Lillis, 2001, 2003, 2006; Lillis and Scott, 2008). Although academic literacies research usually favours a small-scale, ethnographic approach, it still provides a useful set of understandings for the context of my analysis of student writing. In aligning myself with these understandings I am distinguishing the descriptive approach followed in this study from other, deficit views of student writing.

2.6.2 Comparison of ICLE and BAWE

This section contrasts ICLE and BAWE as two corpora which are compiled from the two different theoretical perspectives discussed in 2.6.1: deficit and descriptive approaches. Sections 2.2 – 2.5 indicated that most of what is known about L2 writing is from learner corpora studies and, while these have provided valuable insights into L2 student writing, the findings cannot be directly applied to other forms of NNS academic writing. As Nesi points out:

> although learner corpora provide some insight into the type of tasks language teachers set, they do not represent the type of writing undertaken outside the language classroom. In contrast to language learning tasks, writing for academic or professional purposes usually requires advance preparation, extensive referencing

> to extratextual sources or data, and accommodation to the norms of a particular discourse community. (2008b: 4).

The short argumentative essays comprising learner corpora thus do not adequately represent the multi-genred, disciplinary-specific assessed writing required at undergraduate level. This section compares ICLE, since this is the most extensively-exploited of the available learner corpora, with BAWE, as a recently-created corpus within an under-researched area. The same critiques of ICLE could, however, also be applied to Chinese-specific learner corpora, corpora of IELTS, ToEFL, NMET, CET or other test data[9]. Table 2.1 lists factors specific to the conditions of writing, and Table 2.2 gives factors specific to the student participants in the corpus.

	Learner corpus texts (e.g. ICLE)	**Undergraduate assignments (BAWE)**
Factors specific to the conditions of writing		
(1) Authenticity	Tutors might ask students to produce argumentative essays specifically for the learner corpus.	The writing is naturalistic as it is completed for the external purpose of satisfying course requirements.
(2) Genre	The majority of texts are argumentative essays.	There is a wide range of genres (e.g. case study, essay, reflective writing).
(3) Topic	Essay topics cover a range of accessible, real world topics. Students are unlikely to have any specialist disciplinary knowledge of the area they are writing about. Their answers thus draw on anecdotes and personal experiences.	Students are writing assignments within their chosen discipline. The majority of the writing draws on external sources (though note reflective writing draws on students' own experiences).
(4) Influence of title	Texts are produced from a short choice of titles given to students (see list of IELTS titles in Appendix A). These may promote a dialogic style e.g. soliciting 'your opinion'.	Titles may be provided, with varying degrees of choice, or students may devise their own[10] (see sample in Appendix A).

9 International English Language Testing Service (IELTS) and the Test of English as a Foreign Language (TOEFL) are proficiency tests in academic English which act as gatekeepers for English-speaking HEIs. Gaokao (literally 'tall test') is the entrance test for PRC universities. The College English Test (CET) is a test taken during by PRC students during their university studies in China in order to show progress in English language.

10 Information on whether a title is devised by a student or provided for them (and for the latter whether there was a choice of title) is not given in BAWE contextual data.

(5) Time allowed	Little preparation time is given if writing is conducted in class. Little or no redrafting of essays.	Students can take as much time as they wish on preparation and drafting before producing a final version for assessment. Generally, assignments are set over an extended time period within a term or semester.
(6) Combining writing with reading	No source texts are used although reference books and dictionaries can be used. Horowitz (1986, in Moore and Morton, 2005: 63) describes test writing as 'content-free writing' as it does not display textual plurality that is, there is no citation of other texts. Writing is thus a separate activity from reading.	Writing is 'text-responsible prose' (Leki and Carson 1997: 41) in which students are expected to read first and to cite relevant material (cf. discussion of reading-based writing in Horowitz, 1986; Baba, 2009).
(7) NS support	No NS help is allowed.	Generally, students may use any resources available to them, for example they may ask NS or NNS peers to read and comment on their work.
(8) Length of texts	Texts are short (500-1,000 words) with more at the lower end of this range.	Texts are of variable lengths, ranging from 500 to 10,000 words.
(9) Proficiency of the writing	All writing from 'advanced students' in the cohort is collected (i.e. years 3 and 4 of undergraduate study). The writing is not graded on proficiency or other factors.	Only texts reaching a 'proficient' standard are collected (i.e. scoring at least 60%). Criteria are devised by each department or lecturer and are likely to include the display of discipline-specific knowledge and ideas, engagement with sources, task achievement, as well as linguistic expression.
(10) Paper vs. electronic resources	Students handwrite their essays; the essays are then keyed in for the purpose of corpus compilation.	Only electronically-submitted assignments were accepted for BAWE. A corpus comprising texts written on computer better reflects the reality for most students, since the writing process for each medium requires different cognitive resources (Stapleton, 2010).

Table 2.1 Comparison of learner corpus texts and authentic undergraduate assignments: Factors specific to the conditions of writing

	Learner corpus texts (e.g. ICLE)	Undergraduate assignments (BAWE)
Factors specific to the students		
(1) L1	Each subcorpus is collected from a single L1 in one country. Subcorpora are organized according to students' L1s (e.g. 'SWICLE' is the Swedish subcorpus of ICLE).	Students are from a range of L1s, representing the diversity of UK universities in the early 21st century.
(2) Discipline background	Students are undergraduates in English language/literature	Students are undergraduates in a wide range of disciplines within 'hard' and 'soft', pure and applied areas.
(3) Year groups	Students are 'advanced' level, which in ICLE means they are third or fourth year undergraduates in their home university[11].	Students are from four year groups: undergraduate years 1, 2 and 3, and masters level.
(4) Range of institutions	Writing from an L1 group may be collected from one cohort in a single university in a country.	BAWE assignments come from four UK universities. The additionally-collected 50,000 words come from a broader range of UK universities.
(5) Contributions per student	Each learner contributes one text only.	Students may contribute between 1 and 10 texts[12].
(6) Longitudinality	The corpus is cross-sectional (i.e. collected from different learners at one point in time).	Mix of longitudinal[13] and quasi-longitudinal (i.e. collected at a single point in time but from students of different year groups).
(7) Motivations	Students' motivation for writing the essays is unclear as the texts are written in class but are (presumably) not assessed.	Students are producing writing for assessments which contribute to their degree, so are highly motivated.

Table 2.2 Comparison of learner corpus texts and authentic undergraduate assignments: Factors specific to students

11 'Advanced' level is a wide category. Research using the Common European Framework categories suggests this varies between B2 and C2 level across ICLE, and that different subcorpora vary (Thewissen et al., 2006). Generally, no account is taken in ICLE research of the educational contexts of each country (Tono, 2009).

12 Students taking joint honours degrees may provide up to 20 assignments to the BAWE corpus (10 per discipline).

13 As collection was carried out from 2005-07, student contributors may have submitted assignments at two or more points during this timeframe or may submit assignments from previous years of study, though note that all assignments were written after 2000.

Tables 2.1 and 2.2 point out the distinguishing characteristics of ICLE and BAWE. Of significance are the varied conditions of writing: in contrast with the texts written specifically for inclusion in ICLE, the BAWE corpus texts were produced in a naturalistic setting, that is, they were collected *after* the act of writing. This accords with Tognini-Bonelli's (2001: 55) view of Corpus Linguistics as dealing primarily with language in use which is 'assumed to be genuine communication of people going about their normal business'. In its favour, however, ICLE is a more homogeneous corpus with all texts from the same genre, similar question types, equivalent lengths, and from students who have a similar background. In contrast, the UK undergraduate writing collected by the BAWE project is more diverse and more complex with many more variables such as different genres and assignment lengths. Crucially, BAWE assignments are not graded primarily on linguistic proficiency; indeed, it is unclear how much bearing this has on the grades awarded.

The remainder of this section consists of further discussion of three of my major critiques of ICLE and similarly constructed learner corpora: the influence of titles, the effect of test preparation, and the lack of comparable corpora.

Influence of titles

The first critique discussed here is the influence of titles. Unlike those in the reference corpora of the BNC or LOCNESS or in BAWE, learner corpus essay titles frequently include strong prompts for students to give their opinions with the inclusion of a statement followed by a direct request for the student's opinion. For example:

(1) In the 19th century, Victor Hugo said: 'How sad it is to think that nature is calling out but humanity refuses to pay heed.' Do you think it is still true nowadays?
(2) Some people say that in our modern world, dominated by science technology and industrialization, there is no longer a place for dreaming and imagination. What is your opinion?

The essay titles provided for L2 writers in the context of tests or EFL classes may lead students to believe a dialogic, personal style is acceptable through the use of *you* and the opinion-seeking nature of the titles. Other ICLE titles simple give a statement on a general knowledge topic, with the implication that students should discuss the topic:

(3) The role of censorship in Western society.

A third question type is to provide an opinion on a contentious topic, again with the implication that students will discuss the topic and give their own opinion in their answer:

(4) A man/woman's financial reward should be commensurate with their contribution to the society they live in.

Titles tend to be on topics with a 'for' and 'against' argument, encouraging similar answer formats of lists of advantages and disadvantages followed by the writer's opinion. In this context, high use of personal pronouns seems entirely appropriate when writers are asked to give their opinion on a topic based on their general knowledge alone.

In contrast to those for learner corpora, titles for undergraduate assignments in BAWE are often, complex and lengthy and contain reference to specific information from the course.

(5) 'If a realistic medical jurisprudence is to develop, judges must extend their focus beyond rights and duties and confront the fundamental issue of resources.' Discuss. (Year 2 Law)
(6) Using your hypothetical case study EITHER explain the relevance of different motivational theories to OR explore the role of communication in the situation described. (Year 1 Agriculture)

Further examples of BAWE assignment titles are provided in Appendix A.

Effect of test preparation

A second criticism is the effect of test preparation on the writing produced by students in learner corpora. The preparation and teaching experienced by L1 and L2 student writers prior to essay writing is very different. NNSs are likely to be far more accustomed than NSs to writing short argumentative essays on general knowledge topics, since this type of writing is common in high stakes English language proficiency tests (e.g. the National Matriculation English Test (NMET), taken by Chinese school-leavers, or part 2 of the IELTS writing paper). The screenshot in Figure 2.2 gives an example of the type of writing required for the 'composition' section of NMET.

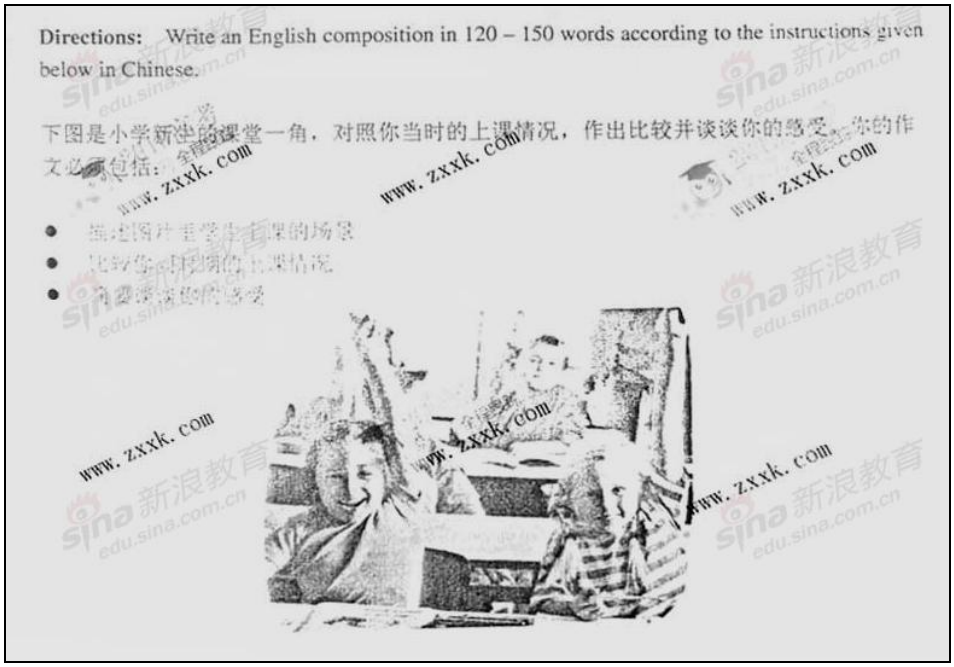


Figure 2.2 Sample instructions for English composition paper[14]

The 'directions' on the screenshot are:

> Write an English composition in 120-150 words according to the instructions given below in Chinese.
>
> (translation): The following is a picture of a lesson from new students in a primary school. Please compare it with your own experience in your classroom and comment on it.

NNS students' experience of writing in English prior to university is often limited to these very short texts, meaning that when given learner corpus essay prompts, they may tend to produce the same kind of limited responses.

Lack of comparable reference corpora

The final critique I make is the distinct lack of comparable reference corpora for research into ICLE, with many studies comparing student writing with professional academic writing. Frequently the 'expert' reference corpus used is the academic writing component of the British National Corpus; the justification for this is frequently that research articles represent good academic writing and as such form a 'target' for student writers. However, ICLE and the BNC differ widely in, for example, the professional level of the authors, the purpose of the texts, and the length and topic of texts. While the ICLE corpora comprise short pieces of

14 Source: http://shiti.edu.sina.com.cn/paper/47/20/32047/c_p.php.

argumentative writing on general knowledge topics, the BNC-academic contains long, discipline-specific writing. Although ICLE students are permitted to use language reference books they do not have to carry out any research in order to write their essays. Usually, in learner corpus research the subset of untimed essays are selected for comparison with (similarly 'untimed') expert writing; yet it is unclear how much time students are likely to spend on a short argumentative essay which is not written for the purpose of assessment. It is also unlikely that these essays have undergone a process of redrafting over an extended period of time in the way that published articles do; moreover ICLE essays will not have had the extensive input of reviewers and editors.

Although learner corpora are routinely compared with expert corpora for 'overuse' and 'underuse' of particular features, in some cases these features could be appropriate in the learners' context of short, argumentative essays. Gilquin and Paquot (2008) argue that evidence from learner corpora points to NNSs' 'overuse' of the chunks *for example* and *for instance* compared to the use of these features by 'experts'. However, the individual texts in ICLE and the BNC-academic are of very different lengths with an average of 500 words in the former and 2000 in the latter. If each 500-word essay introduced just a few examples using these chunks, the result might be 'overuse' in the whole corpus compared with the BNC texts. Thus, the 'overuse' remarked on by Gilquin and Paquot could be due to the compressed staging of an argument within a short essay, such that an exemplification stage is reached far sooner than in a 2000-word article sample.

A further issue to raise in considering the use of expert writing is that while research articles could be said to constitute a single genre, student undergraduate assignments, in the UK at least, are currently undergoing rapid expansion in the range of genres expected. Some of these assignments are becoming less similar to research articles; for example the increased requirement that students reflect on their work results in the inclusion of very personal writing with a large number of personal pronouns and mental verbs; in the same way the use of real life writing such as letters or CVs as an assignment task is also further removed from the

professional academic writer's task. A more appropriate reference might instead be academic blogs or professional letters.

Since comparing L2 writing with expert writing is clearly not comparing similar genres, comparing L2 writing with texts from L1 writers of a similar age and educational level might seem a better choice. The LOCNESS corpus of British and American university students' argumentative writing was compiled at the Université Catholique de Louvain specifically for comparison with the short, argumentative writing in ICLE (e.g. studies by Gilquin and Paquot, 2007; Guo, 2006; Ringbom, 1998; Wiktorsson, 2003). However, the two corpora are not exact equivalents. Many of the LOCNESS essays are longer than the standard 500-word ICLE essays as the former includes university discipline-specific writing which makes use of background reading. Moreover, the writers of LOCNESS texts are motivated to complete the writing task for the instrumental reason of gaining a good mark in their A-level or university essay; for ICLE writers, motivation is limited to performing well in the class since the texts are not for any broader purpose (at least as far as the students are concerned). The list of titles provided by the different institutions contributing essays to LOCNESS indicates that these seldom ask a question in the title, rarely include a pronoun, and tend to be short, sometimes consisting of a single noun phrase (e.g. 'Euthanasia', 'Fox hunting', 'Water pollution'). While there are many points on which ICLE and LOCNESS are similar, for example the general knowledge nature of the topics and the requirement to write an argumentative essay, this difference in titles may partially account for the variation in dialogic style and in pronoun use by the student groups.

One way of overcoming the deficiencies in the reference corpora is to employ more than one reference corpus. For example, Chen (2009) compared Chinese students' writing with both NS students and professional academics (4.3.2 contains further discussion of reference corpora).

Summary

Section 2.6 has argued that learner corpora, while offering valuable insights into NNS writing, consist of short, decontextualised texts from a single genre and so are very different

to undergraduate assessed writing. This study seeks to improve on this means of describing student writing by investigating undergraduate student writing produced for the purpose of assessment and which is written for discipline lecturers rather than ELT specialists. Both Chinese and British students are writing for the same purpose and, as far as it is possible to ascertain, receive the same input in terms of teaching through lectures and seminars, feedback on their writing and exposure to written academic language through recommended readings.

2.7 The research questions

The beginning of this chapter presented review questions which were posed of the available research literature on Chinese students' academic writing; this section discusses the research questions arising from these initial questions in the light of findings from the literature. Many of the studies reviewed in this chapter focus on lexical chunks within student writing (e.g. Chen and Baker, 2010; Gilquin and Paquot, 2007; Paquot, 2010; Hyland, 2008a,b; Lee and Chen, 2009; Li and Schmitt, 2008; Thompson, 2009; Wiktorsson, 2003). Lexical chunks are a common feature of analyses of academic writing as they represent preferred, conventionalized ways of expressing meaning and are regarded as indicators of competent language use (e.g. Biber and Barbieri, 2007; Cortes, 2004; Hyland, 2008b). Central to this thesis, therefore, is the exploration of Chinese and English students' use of lexical chunks through a corpus investigation of authentic university undergraduate assignments. From this data, similarities and differences will be drawn out between the two student groups overall, between writing from the first two years and the final year of undergraduate study for Chinese and English students, and between writing in three selected disciplines.

In the study, three research questions (RQs) are specifically addressed.

RQ 1: What are the distinguishing characteristics of writing in English in a corpus of Chinese undergraduates' assignments in the UK?

Among the characteristics identified in the literature are the high use of particular lexical items and chunks including informal or 'speech-like' items, particular connectors, and the high use of personal pronouns, and it is anticipated that the Chinese students' texts from BAWE might show some evidence of these characteristics. As this study is corpus-driven, employing keyword analysis to explore the data (see 4.3.2 for discussion of keyword analysis), this allows additional unexpected findings to emerge which may extend or contradict those found in the research literature. RQ1 is explored within Chapter 5.

RQ 2: Are there any variations in the characteristics identified in this study between years 1/2 and year 3?

This research question focuses on variations in the writing produced by Chinese students at different stages of undergraduate study. First and second year students' writing will be compared to year 3 students; though it should be recognized that since these groups contain (for the most part) different individual students, large claims cannot be made as to *progress* but instead tentative suggestions can be made as to *variation* across year groups. The writing of year 3 students of either L1 may have varied due to factors such as the feedback received from tutors, reading carried out within their discipline, input from lectures and seminars, and general acculturation into both UK academia and into their discipline. RQ2 is the focus of Chapter 6.

RQ 3: In what ways do disciplines affect the identified characteristics of Chinese undergraduate writing in English?

This question explores the three most popular disciplines for Chinese students (Biology, Economics and Engineering) and considers how each of these differs from undergraduate writing overall. For example, in what ways does undergraduate writing in Engineering differ from all undergraduate writing? The scope of Chapter 7 is limited to consideration of previously-identified features: differences in the use of *we* and *I,* variation in the use of connectors across disciplines, and differences in the use of visuals and lists. The multimodal aspect of assignments is explored through both Corpus Linguistics and through a whole text

investigation of pairs of texts from Chinese and English students in the same disciplines (Biology, Economics, Engineering) and answering the same assignment question.

2.8 Chapter summary

This chapter has surveyed the literature on Chinese students' undergraduate writing in the UK; due to the paucity of recent corpus studies in this area this narrow focus was broadened to consider studies beyond undergraduate level. Much of the available literature concerns studies of learner corpora as in recent decades these have provided the main datasets for investigations into Chinese students' writing, and NNS writing more widely. Findings from these are generally agreed on NNS' high use of particular lexical items and chunks, including informal or 'speech-like' items and connectors, and high use of first and second person pronouns. However, the texts in learner corpora are very different to those at undergraduate level and it is unclear how far the findings from the former can be extended to the latter. Learner corpora consist of 500-word, argumentative essays on a limited range of topics, do not require background reading or research, and analysis of the texts is limited to features of linguistic proficiency. In contrast, undergraduate assessed writing covers a wide variety of disciplines and fields within disciplines. The review of the available literature has confirmed that the area of Chinese students' undergraduate writing is an under-explored field. The issues of variation across year groups and writing in different disciplines at undergraduate level were also identified as gaps in the current literature and as therefore worth pursuing in this study.

The data in this study consists of authentic undergraduate writing produced in a natural setting rather than data produced for a test or for purposes of analysis. The comparison dataset of L1 English undergraduate students' writing has the same external conditions of writing and collection, as far as possible, rather than being a corpus compiled from different universities and on different topics (as with ICLE and LOCNESS). This study contains quasi-longitudinal data, allowing comparisons between different year groups to be made. While these comparisons are not of the same students, and hence are termed 'quasi-longitudinal', they give an indication of the way in which characteristics in a cohort's writing may vary over

time. Finally, the data is taken from a range of disciplines, allowing some comparisons to be made and tentative suggestions to be given as to how writing may vary across disciplinary groupings.

The ensuing research questions for the study focus on each of these three areas and are addressed in Chapters 5, 6, and 7. In the intervening chapters, the characteristics and identification of lexical chunks is discussed (Chapter 3) and the research methodologies used in the study, namely Corpus Linguistics and whole text analyses, are explored (Chapter 4).

CHAPTER 3 ANALYZING STUDENT WRITING THROUGH A FOCUS ON LEXICAL CHUNKS

3.1 Introduction

Chapter 2 reviewed findings on Chinese students' writing, situating these within research into NNS writing more broadly, and also discussed studies of NS writing in UK universities. A large number of these studies focus on lexical chunks in student writing, using this phenomenon as a means of measuring differences in writing between NS and NNS, and/or accounting for 'naturalness' in writing (e.g. Chen, 2009; Chen and Baker, 2010; Gilquin and Paquot, 2007; Granger, 1998; Paquot, 2010; Hyland, 2008a,b; Lee and Chen, 2009; Li and Schmitt, 2009; Thompson, 2009; Wiktorsson, 2003). Much of this literature directly states that lexical chunks are highly significant in any understanding of student writing. For example, Li and Schmitt (2009: 85) describe lexical chunks as the 'defining markers of fluent writing', and Hyland (2008b: 4) states that chunks are both 'central to the creation of academic discourse' and 'a key factor in successful language learning'.

'Lexical chunk' or simply 'chunk' is used in this thesis as a broad term referring to any wordforms which can be classed as a unit, whether through frequency, internal coherence, or other linguistic features. Recent experimental studies have provided evidence to support the belief that psychologically-salient chunks are easier to recall and produce, and are thus probably stored as wholes in both the NS and NNS language user's mental lexicon rather than being compiled item by item (e.g. Conklin and Schmitt, 2007; Ellis et al., 2008; Erman, 2007; Jiang and Nekrasova, 2007; Wiktorsson, 2000). It has always been difficult, however, for theorists to agree on the proportion of the lexicon that is stored in chunks. Much depends on the genre considered (e.g. spoken or written; academic or non-academic; formal or informal); the type of language classified as a chunk (e.g. fixed idioms or open frames); and the means by which chunks are found (e.g. through a top-down list of pre-determined chunks or through bottom-up, computationally-extracted searches). Estimates for the extent of language remembered as chunks for NSs of English range from 80% (Altenberg, 1998); 32%

(Foster, 2001); 28% (Biber et al., 2004) to 4-5% (Moon, 1998a), though most researchers would agree with Pawley and Syder's (1983: 205) earlier assertion that the NS lexicon contains 'some several hundred thousand sequences'. However, the situation is likely to be different for NNSs learning through a formal school setting and with little exposure to English. For example the Chinese students in this study learned English through detailed analysis of sentence-level lexis and grammar; such an analytic learning style is probably more conducive to a word by word approach to language use than the high use of lexical chunks acquired by NSs.

In addition to a primary focus on lexical chunks, this study makes use of overall text characteristics comprising mean assignment length (measured in tokens), mean sentence length (measured in tokens) and mean word length (measured in characters per wordform). These features help to build a 'linguistic profile' of each dataset, allowing broad comparisons to be made (measures of text characteristics are discussed in 4.3.1). Lexical chunks, however, form the main part of the discussion and are explored through the extraction of 'keyness' (i.e. linguistic items which are statistically more frequent in one corpus when compared to another), and through comparison of the most frequent chunks across the student corpora. Some findings are extracted at the wordform level (as 'keywords') for completeness, though the ensuing discussion concerns the patterning of these words within the surrounding text.

The remainder of this chapter focuses on the description and analysis of lexical chunks as the main starting point for the exploration of the student writing in this study. Section 3.2 gives an overview of selected language theories within the area of phraseology; and 3.3 extends Hoey's lexical priming theory to consider the development of L2 writing. Section 3.4 situates the methods for ascribing characteristics to chunks and extracting them from data before the definitions and methods used in this thesis for lexical chunks are discussed in 3.5. Finally, once extracted and identified, it is useful to classify lexical chunks both structurally and functionally in order to facilitate comparisons of usage, and the taxonomies utilized in the study are explored in 3.6.

3.2 Relating lexical chunks to student writing

This section discusses selected theories relating to lexical chunks, namely Miller's (1956) information processing theory; Hopper's (1987, 1998) emergent grammar; Sinclair's (1991) idiom principle; Hunston and Francis' (2000) pattern grammar; and Wray's (2002, 2008) Needs-only analysis. These theories of chunks then provide a basis for my extension of Hoey's lexical priming to a consideration of NNS academic writing.

Theories explaining lexical chunks

I begin this section with a theory which is not confined to language but embraces wider cognition, and paves the way for the subsequent theories on lexical chunks. George Miller's (1956) information processing theory suggests that humans process information of any kind in meaningful units or 'chunks'. Miller's notion of a chunk as 'any highly familiar unit' (p.82) has been used extensively in fields as diverse as remembering sequences of moves as chunks in chess psychology (Chase and Simon, 1973); producing sections within a musical score (Jackendoff, 1988); and 'chunking' activities within Neuro-Linguistic Programming (Craft, 2001). What makes these 'familiar units' particularly beneficial for learning is the proposition that 'it is possible to expand the total amount of information by packing more and more information into one chunk' (Howard, 1983: 104). This explains how the chess expert is able to memorize sequences of moves considerably longer than those of the novice player, and the skilled language user commands a repertoire of lengthy chunks of language while the lower-level user struggles with short sequences. Schmitt and McCarthy (1997: 230) draw on Crick (1979) to similarly argue that the mind 'uses an abundant resource (memory to store prefabricated chunks of language) to compensate for a limited one (processing capacity)'. These twin notions of the vast storage capacity for lexis and the limited processing capacity of the human brain explain the usefulness of lexical chunks in language acquisition.

While Miller's proposal of information being stored in chunks provides the basis for a psychological explanation for language fluency, it was Hopper's (1987, 1998) radical approach to grammar that heralded the beginning of new linguistic theories. Hopper's work

on emergent grammar has at its core the premise that previous notions of grammar (which were dominated by Chomsky's, 1957, Universal Grammar) were abstract systems which viewed grammar as a set of rules existing independently to language users. Hopper (1987: 3) inverts this, suggesting instead that grammar emerges from the language used by people, thereby it is 'always in a process but never arriving, and therefore emergent'; there is thus no abstract grammar to analyze separately from language in use, but 'only 'grammaticisation' (p.7). Hopper's notion of 'grammar' is the outward manifestation of common patterns, or what he terms 'routines'; use of these results in separate grammars for each individual language user. This radical move from an abstract system to the idea of individual, emergent grammars distinguishes Hopper's work from that of previous formalist grammarians, and underpins the work of later linguists (e.g. Hoey, 2005; Hunston and Francis, 2000; Sinclair, 1991).

Also focusing on the individual's choices in language is Sinclair's (1991) development of the idiom and open choice principles which combine psychological and linguistic views to produce a model of language production and comprehension. In a few succinct pages, Sinclair outlines his theory of how words co-occur, arguing that two principles are required to explain language use. The idiom principle states that:

> a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analyzable into segments (Sinclair, 1991: 110).

This principle is, Sinclair argues, the major way in which language is produced and understood, however 'its importance has been largely neglected' (p.111). Where there is no pre-constructed chunk to draw on, the language user is forced to follow the open principle and construct an utterance from scratch, drawing on their (conscious or unconscious) knowledge of grammar. Almost all grammatical models are based on the open principle, combining individual wordforms with grammatical rules in a 'slot and filler' approach to language. The two outlined principles cannot operate simultaneously, meaning that at each choice boundary, either the idiom principle or the open principle is selected. Within both

reception and production, the user first follows the idiom principle, that is, employs a single choice chunk, thus:

> The first mode to be applied is the idiom principle, since most of the text will be interpretable by this principle. Whenever there is good reason, the interpretive process switches to the open-choice principle, and quickly back again (p.114).

Sinclair does not expand at length on why the idiom principle is preferable, suggesting that it requires less effort to select a prefabricated chunk than to construct language anew and that real time communication also plays a role. However, he argues that '[h]owever it (the idiom principle) arises, it has been relegated to an inferior position in most current linguistics, because it does not fit the open-choice model' (p.110). From this 'inferior position' at the time Sinclair was writing (1991) has sprung a wealth of theory and literature surrounding the importance of lexical chunks. Two theories drawing on the idiom principle are considered here: the pattern grammar of Hunston and Francis (2000), and the Needs-only analysis of Wray (2002, 2008).

Hunston and Francis (2000: 102) draw heavily on Sinclair's view of single choices in their theory of pattern grammar. This theory states that lexical items are associated with various grammatical patterns; for example, *matter* when used as a noun is usually preceded by *a* and followed by *of* plus a clause beginning with an *-ing* form (*a matter of learning.., a matter of being able to*…) (p.2). The patterns of a word are, then, 'all the words and structures which are regularly associated with the word and which contribute to its meanings' (p.37). In the extensive examples provided by Hunston and Francis, as in the related Collins Cobuild dictionary, patterns are given with parts of speech in place of the possible lexical instantiations. For example the (straightforward) pattern:

V n

can be exemplified as:

she ate an apple.

and the (more specific) pattern:

V n *into –ing*

is instantiated as:

He talked her into going out with him.

Other verbs taking the same pattern as *talk* include *frighten, intimidate, panic, scare, terrify, embarrass, shock, shame* (all of which, as Hunston and Francis point out, are connected with negative emotions), and *calm, soothe, relax* (which are not) (p.102). Thus, items within the same semantic grouping may adopt the same pattern structure. Although they rely extensively on parts of speech (or 'word class') to describe patterns, Hunston and Francis adopt the view that structural part-of-speech labels are useful abstractions rather than signalling an external reality:

> the notion of class is just a convenient short-hand: it is easier to say that a word is a count noun than to say it is preceded by 'a' or 'some' and or by other 'open-set' words (2000: 197).

Thus word classes are 'necessary in order to make sense of the huge range of behaviour that words have', and the 'problem' is then to find the right number of word classes to adequately explain lexis without obfuscation (p.197).

Hunston and Francis' research is inspired by corpus research and is fully corpus-driven; indeed, a large corpus is deemed essential for lexico-grammatical patterning to be observable. While concerned with finding and explaining language 'patterns' rather than 'chunks', the 'grammar patterns' they outline are 'in a sense examples of lexical phrases' and '[w]riters on lexical phrases and on grammar patterns… seek to account for some of the same evidence in different ways' (p.14). Patterns are thereby 'a valuable way of finding useful generalisation among a mass of information about individual lexical items' (p.272), and can be viewed as a principled way of accounting for all frequently-occurring language.

Also adhering to Sinclair's idiom principle is Wray's needs-only analysis (2002, 2008), which argues that language is produced and understood in chunks rather than as individual lexical items. Chunks (termed 'formulaic sequences') are meaningful units and are broken down into individual lexical items and wordforms on the basis of need only; where the language user does not need to analyze a sequence into its component parts, it remains intact in the mental lexicon. Hence, the default is for language users to follow the idiom principle and to

use formulaic sequences. In considering how entire communities come to possess largely the same set of formulaic sequences, Wray argues that individuals build up an 'inventory' of sequences which are 'heavily influenced by the current patterns of usage in the speech community' (p.74); thus each time an individual uses an item from their inventory they in turn 'contribute to what others hear most often and therefore store in their own inventories' (p.74). She ascribes the motivation for this as both due to processing constraints and our 'desire to sound like others in the speech community' (p.75) (cf. Wray and Perkins, 2000). Although concentrating on spoken rather than written language, Wray's argument can be equally applied to academic writing: language users employ items read in the writing of others through a wish to write in the preferred academic style. If a majority of prior reading is from ELT textbooks produced in China which blur the distinction between spoken and written, informal and formal, then this conflating of registers will be produced in writing (at least until evidence to the contrary is perceived).

Wray (2008) draws on cognitive and construction grammars to explain the phenomenon of formulaic language. Gries (2008) outlines some of the approaches subsumed under these terms. He acknowledges that while cognitive grammar (e.g. Langacker, 1990) does not specifically account for lexical chunks, it does dispense with a strict demarcation between lexis and grammar (cf. Hopper, 1987; 1998). The 'symbolic units' of cognitive grammar are pairings of any kind of form and meaning or function, such that the more frequently a pairing is encountered by a language user, the more 'entrenched' the unit becomes and the more easily it can be accessed (Gries, 2008:13). Construction grammars (pluralized since there are several versions, e.g. Croft and Cruse, 2004; Fillmore et al., 1988, in Speelman et al., 2009) follow broadly the same lines but with the additional requirement that a symbolic unit or construction requires non-compositionality; in other words, these units cannot be understood from their component parts alone but are used as a whole chunk. Both cognitive and construction grammars emphasize the importance of frequency of use, making them highly compatible with a performance-based view of lexical chunks.

Unlike the linguistic theories discussed so far in this section, Hoey's (2005) lexical priming does not seek to specifically discuss or explain lexical chunks. Instead, Hoey provides a theory of language beginning with the word. However, this theory can also be applied to chunks and, as Hoey (2009) himself argues, is compatible with existing theories which focus on chunks. Hoey (2005: 8) puts forward a convincingly argued case that each and every word is 'primed' for language users, meaning that we acquire knowledge of the word's collocations, colligations, semantic associations, textual positioning, and other features pertaining to its use. This process is gradual and ongoing in that we continue acquiring new words and adapting our primings, and the theory seeks to account for the 'naturalness' of words and sequences in use within a particular discourse community:

> As a word is acquired through encounters with it in speech and writing, it becomes cumulatively loaded with the contexts and co-texts in which it is encountered, and our knowledge of it includes the fact that it co-occurs with certain other words in certain kinds of context (Hoey, 2005: 8).

Note that lexical priming is a property of the *person* rather than the word, though Hoey's writing at times contains some slippage in this distinction. In an echo of Hopper's emergent grammar (1987, 1998), Hoey states that the resulting collective 'primings' (p.7), or knowledge about words, form the collective natural use of language. These primings are not permanently fixed, but at times 'crack' (p.11) due to conflict with other people's primings or undergo change through a more gradual 'drift' (p.9) as differences in meaning and use are encountered. Hoey's description of primings evokes a (metaphorical) organic structure in that 'cracked' primings can 'heal' (p.11) or 'mend' (p.11) through, for example, the rejection of either the original or the new priming, or by assigning the conflicting meaning to a different context.

Hoey's rationale for beginning with the word, rather than the lexical item or lexical chunk, is that this is a 'convenient starting point' (p.158); however, this seems an insubstantial basis for an otherwise cogently-argued theory of language. The dominant focus on words (as individual orthographic wordforms rather than combining to form lexical items) is sustained throughout the book (though there is some slippage in the use of 'word' and references to 'vocabulary'). Hoey suggests that words are acquired first and are primed, then these

primings combine to form lexical items, collocations and formulaic sequences, as a 'second phase in the priming' (p.8). This process of combining words is termed 'nesting' (p.8) and results in chunks which are more than the sum of their parts, hence 'the product of a priming becomes itself primed in ways that do not apply to the individual words making up the combination' (p.8). Nested lexical items and lexical chunks then, like words, 'become loaded with the context and co-texts in which they occur' (p. 8).

Lexical priming and compatibility with other theories

Hoey articulates and extends ideas that have been in circulation since Hopper's publication on emergent grammar in 1987. His contribution is in part the bringing together of theories by, for example, Sinclair, Wray, and Hunston and Francis. For example, lexical priming theory echoes Wray's discussion of how people use the same set or 'inventory' of sequences which sound right because they have been heard so frequently (Wray, 2002: 74, and discussed earlier in this section). Wray's (2002) description of acquiring formulaic sequences from other people's speech coheres with Hoey's discussion of the harmonisation[15] of primings through language encountered from sources such as television, one-to-one conversation and education. Chinese undergraduate students enter year 1 with an 'inventory' (Wray) or set of 'primings' (Hoey) from their experiences of English texts in China, and by year 3 have reconfigured their inventory to include more recent primings from UK study. However, Hoey deviates from Wray's views in terms of the emphasis placed on chunks. Whereas Wray's (2002) needs-only analysis views formulaic sequences as primary, to be broken down on a 'needs only' basis, Hoey (2009: 35) begins with the word and views chunks as the endpoint of a series of 'priming encounters'. The two theories are broadly compatible, but with the different starting point of Hoey meaning that words, rather than chunks, are foregrounded.

In a similar way, Hoey's ideas on priming are, he argues, compatible with Sinclair's (1991) idiom principle. Hoey (2009) provides a detailed explanation of how the chunk *dry up*, which it might be assumed is formed through the idiom principle, can be seen as the product of various layers of primings (such as the collocation of *dry* and *up*, the colligation (or

15 Throughout this thesis the 'ize/ization' ending is preferred, except for 'harmonise/harmonisation' as these are Hoey's specialized terms.

grammatical collocation) with intransitivity, the textual semantic association with a problem as in Hoey's example 'when the supply of school-leavers starts to dry up').

Hoey (2009) also suggests that there is no conflict between Hunston and Francis' (2000) pattern grammar and lexical priming, only a difference in focus since the former is concerned with providing grammatical abstractions for lexis whereas the latter begins from the individual's internal grammar. Pattern grammar states that:

> A pattern is a phraseology frequently associated with (a sense of) a word, particularly in terms of the prepositions, groups and clauses that follow the word (Hunston and Francis, 2000: 3).

Thus, the patterning surrounds the word and is the same for all language users. In contrast, for lexical priming:

> what we think of as grammar is the product of the accumulation of all the lexical primings of an individual's lifetime. As we collect and associate collocational primings, we create semantic associations and colligations (and grammatical category primings). These nest and combine and give rise to an incomplete, inconsistent and leaky, but nevertheless workable, grammatical system (or systems) (Hoey, 2005: 159-160).

In other words, there is no external system of 'grammar', only a collection of individual grammars which are formed over time and are constantly changing. For Hoey (2009: 46), then, pattern grammar 'produces the generalisations that account for majority practice' and lexical priming provides an explanation as to why these generalisations do not accord with the practice of each individual language user.

I now draw out some shared features from the theories of Hopper, Sinclair, Hunston and Francis, Wray, and Hoey concerning the nature and primacy of lexical chunks. First, linguistic theories pertaining to lexical chunks are rooted in real data (corpus or otherwise), rather than abstract descriptions of how language works based on linguists' contrived data; this distinguishes them from previous generative grammars. Second, lexis is primary, meaning that any discussion of language should be 'bottom-up' rather than 'top-down'. Grammatical 'rules' are thereby seen as abstractions, serving to flesh out the 'phraseological

skeleton' (Philip, 2008: 97). Crucially, lexis and grammar can be viewed as a single entity (the 'lexicogrammar') rather than as separate elements. Third, lexical chunks are at the core of language use. It is much easier for language users to employ chunks of language than to construct language anew from individual lexical items. Proficient language users possess large numbers of lexical chunks, many of which contain variable parts, and they are constantly adding to their repertoire. Effective use of a chunk entails knowledge of the circumstances in which it can be used, that is, an awareness of prior uses and patternings. Finally, lexical chunks are not fixed but are in a constant state of flux: re-using a chunk reinforces its patternings for both producer and receiver and ensures its continued use in the future.

Towards a theory of L2 writing

This section expands Hoey's lexical priming theory to consider L2 language learning, and in particular, how Chinese undergraduate students in the UK have been primed by their previous exposure to written English. Lexical priming is taken as the starting point as this theory draws together previous notions of emergentism, the primacy of lexis, and the effect of previous language learning experience.

Following the lexical priming theory, it might appear that the more times a linguistic item is encountered, the greater the chance of it being primed for a particular individual. In addition to these multiple encounters from the language of others is the phenomenon of individuals producing sequences and effectively priming themselves, either in the presence of others or alone as 'private speech' (Vygotsky, 1934/1986, in Lantolf and Thorne, 2006). From this, however, it should not be concluded that lexical priming views language learning as simply a behaviourist process driven by the frequency with which linguistic items are encountered; it is more that repeated encounters have their part to play. 'Noticing' (Schmidt, 1990) is also key in determining which items are primed for particular users. Thus, supposing it were possible to construct a corpus for a particular individual's receptive and productive language, that is, capturing everything the person says, writes, hears or reads (both publicly and privately), we would still not be able to predict which primings would be formed most strongly as we would not know the significance of each linguistic occurrence. For example, pertinent

factors include whether instantiations are from ‘respected’ sources, or whether chunks are encountered when the individual is most alert. The researcher would also fail to capture the effect of previously-acquired languages and primings and the use of translation-equivalent chunks in their prior language use.

The adult L2 user might learn through a dominantly analytic approach or through a holistic one. The approach adopted is likely to depend on the circumstances of learning the L2 (e.g. as a foreign language within an L1 environment or as a second language in the target language context) and on the method of learning (e.g. Grammar Translation with its inherent focus on lexical equivalents for translation, or a communicative approach with a concentration on learning chunks of natural language). The literate adult learner who has learned through an analytical method will *expect* lexical chunks to break down into smaller linguistic components (lexical items and wordforms). Since NNSs learning through a dominantly Grammar Translation approach have acquired their primings through the eye rather than the ear, the orthographic wordform is particularly salient (cf. Bassetti’s, 2005, 2009, work on the effect of interword spacing in English and for Chinese characters written in Romanized script [pinyin]).

Following the initial primings, the language user becomes sensitive to how their primings are shared with or distinct to other people’s, and begins to abstract from their primings. No-one’s primings are exactly the same since individuals have different linguistic histories: each person is exposed to a unique array of conversations, stories, people, education, and other influences, though there is enough in common for a shared language to exist. Hoey argues that commonality in language use is achieved via the ‘harmonising’ of language (p.11) through education or self-reflexivity wherein an individual’s primings become similar to those of the majority. A corollary of Hoey’s theory, then, is that the greater an individual’s exposure to the language of the discourse community, the greater the degree of harmonisation of primings. Chinese students are primarily exposed to sentence-level analysis, intensive reading texts, and the short essays used for NMET practice, and exposure to academic English texts is limited. However, on moving from their home country to study in the UK,

Chinese students encounter a different set of primings (the same is true of UK-based students, though to a far lesser extent, as the transition from A-levels in the UK to university writing is not as marked). For Chinese students, primings can drift in a gradual process of learning the preferred discourse of both the academy and of their discipline, or crack due to a sudden clash between their usage and the dominant one of textbooks and lecturers.

For Hoey, we are all developing users and will continue to acquire the lexical chunks pertaining to a new disciplinary area. Every time a student has a conversation with a new person within a discipline, or reads a new research article, they are reinforcing previously-acquired lexical chunks, and are perhaps both acquiring new chunks and reassigning some existing ones. Since lexical items and chunks are primed for language users at each new encounter, there is no end point to the process of acquiring language. The distinction between NS and NNS is reduced since:

> if each person constructs their language out of the primings acquired from a unique set of data, there can be no right or wrong in language (and no absolute distinction between native and non-native speaker, though the latter will have acquired their primings by strikingly different routes) (Hoey, 2005: 181).

These ‘strikingly different routes’ refer to L1 acquisition through a largely non-rule-based process of exposure to language. The L2 route to acquisition, in contrast, may include a focus on grammatical explanations, providing a shortcut to competent usage. However, this egalitarian view of parity between NS and NNS language use seems at odds with the reality of hierarchical structures in academia. A lack of ‘special status’ for any individual’s primings does not hold true in situations where one participant is of higher status than the other. It is likely that students will be primed more readily by academic textbooks and by lectures than by their student peers in terms of academic language. Moreover, the primings of a group of higher-status individuals (e.g. lecturers in a university department) may dominate over those of a lower-status group (e.g. the undergraduate students in the department) within the realm of the academic discipline. In other respects, Hoey’s views accord with an academic literacies view of the gatekeeping function of the dominant discourse in that:

> examinations also seek to ensure that only the examinees whose primings harmonise with those already in positions of influence or power are able to take up positions of influence or power themselves (Hoey, 2005: 182).

Adopting the accepted discourse is thus the route to academic success. However, for NNS students, the transition to UK university study is likely to involve a significant discord with previously-acquired primings.

This section has considered a range of theories relating to lexical chunks, extracting overlapping principles from these, and expanding Hoey's theory of lexical priming to a broader consideration of how both NS and NNS students continue learning language. The next section discusses the characteristics of chunks which can then be used for identification and categorization.

3.3 Identifying lexical chunks

Wray (2008: 93) discusses an inherent circularity in identifying formulaic language, since 'you cannot reliably identify something unless you can define it', yet in order to define it, you must have some examples to study. Any definition of what constitutes a chunk is therefore bound up with the choice of method for identification; for example defining a chunk by its frequency of occurrence leads to a computational method of identification (excepting very small samples where counts can be manual); and a definition based on semantic coherence is likely to rely on intuition to some degree. Some phraseologists precisely specify the characteristics for lexical chunks, depending on their focus (e.g. for Moon, 1998a, 'Fixed Expressions and Idioms' (FEIs) must fulfil exact requirements of lexicogrammatical fixedness, institutionalization and non-compositionality), while others suggest that chunks can satisfy just some of the features in a long list (e.g. the list of diagnostic criteria for assessing intuitive judgments of formulaic language provided in Wray and Namba, 2003). This section thus begins with a discussion of some commonly-used characteristics of lexical chunks, before turning to means of identifying chunks. The aim of the section is to illustrate the range of rationales and the diversity of approaches in current taxonomies, before discussing the computational extraction of chunks employed in this study (discussed in 3.5).

3.3.1 Characteristics of lexical chunks

This section discusses the main characteristics which have been used by a range of researchers in determining what does and what does not constitute a lexical chunk (Wray, 2002, provides an in-depth discussion of different approaches). The first feature relates purely to spoken language, while the remainder can refer equally to speech and writing.

Phonological features

Language produced as a chunk is now believed to be uttered more rapidly than is the case for non-chunked language, without internal pausing, and within a single intonation contour (Erman, 2007; Foster et al., 2000; Wood, 2004, 2007). In a study measuring dysfluencies in speech, Foster et al. (2000: 355) point out that 'the more proficient speaker… [is] the person who can keep track of more complex micro-units', that is, proficiency is linked to the rapid accessing of chunks and thus fluency in production. Similarly, Wood (2007) claims in a study of three Chinese learners of English that chunks helped in the learners' development of fluent speech, as measured by an increased speech rate, decreased number and length of pauses and increased length of speech 'runs' (i.e. fluent speech between pauses).

Institutionalization

Moon (1998a) refers to the procedure in which lexical items become 'entrenched' within the language as familiar expressions, giving the examples of the routine expressions *many happy returns* and *mind the gap*. Erman (2007: 33) describes this as a process of becoming conventionalized, and refers to the 'restricted exchangeability' of these chunks wherein an item in a chunk cannot be substituted without some change in meaning, function or idiomaticity. Moon (1998a: 7) describes how institutionalization can be viewed as a quantitative feature within corpus studies, thus by virtue of being frequent, a chunk could be described as institutionalized.

Lexicogrammatical fixedness

'Fixedness' is a key part of Moon's (1998a) definition of FEIs. Giving the examples of *kith and kin* and *call the shots* she suggests that a 'formal rigidity… implies some degree of lexicogrammatical defectiveness' (p.7) as these expressions are syntactically unchanging.

Wray and Perkins (2000: 11) point out that once stored in the lexicon as a formulaic sequence, fixed expressions can then become 'archaisms' which 'survive language change', for example *if I were* has survived the demise of the subjunctive in English and can therefore be described as 'fixed'.

Semantic non-compositionality/Idiomaticity

Granger and Paquot (2008: 31) define a chunk as non-compositional when 'its global meaning is different from the sum of its individual parts'; that is, its full meaning cannot be adequately understood from breaking the chunk down into individual lexical items, analyzing the meanings of each item and putting these meanings back together. Compositionality is often described in terms of a cline from the fully compositional, such as *open the window*, to the fully non-compositional, such as the idiomatic *raining cats and dogs* (cf. Moon, 1998a). Thus, while the meaning of the former example can be built up or composed from the meaning of individual lexical items, the latter is known as a whole and is likely to be segmented and re-composed only for creative or comic effect (e.g. *it's raining tabbies and chihuahuas*).

Frequency of occurrence vs. semantic unity

The previous features would each allow a single instance of an item in a text or corpus to be termed a lexical chunk. In contrast, the criterion of frequency entails multiple occurrences for an item to qualify and is the primary defining feature of chunks known variously as 'clusters' (e.g. Scott, 2010), 'n-grams' (e.g. Milton, 1999), and 'lexical bundles' (e.g. Biber et al., 1999). In a corpus search, these require parameters to be set for the length of a sequence, threshold for minimum frequency, and the minimum texts for dispersion. For example, for Biber et al. (1999), four-word lexical bundles have to occur ten or more times in a corpus and in a minimum of five texts per register in order to qualify as bundles. These parameters help to avoid idiosyncrasies and also repetitions due to localized topics (Figure 4.1 gives a fuller range of different researcher's parameters). While corpus linguistic software packages can quickly produce a list of recurring contiguous chunks, these can cross structural boundaries and may feel psychologically invalid as whole units (e.g. *way in which the*). Frequently-occurring chunks thus may or may not be 'semantically whole'.

In contrast, semantic unity is the notion that a chunk has coherence, that is, it feels 'complete' to the language user. Chunks are recalled as 'wholes' rather than built up from individual lexical items and can be viewed as simply 'big words' (Ellis, 1997: 130). Sequences have either been acquired by the user as wholes, or have been initially constructed from individual words and then stored and retrieved as formulaic sequences (Wray and Perkins, 2000). Semantic unity is central to the view of chunks as psychologically whole items, and is often in opposition to the notion of frequency as a central defining characteristic. For example, a lexical chunk occurring once only in a corpus (a 'hapax legomenon') might be semantically 'whole' but would not be captured through frequency counts. Conversely, a chunk can occur frequently but not feel semantically 'whole' (e.g. in my data the chunk *that there is a* is a frequently-occurring chunk).

Contiguity

A chunk is described as 'contiguous' if the individual linguistic items (whether wordforms or named parts of speech) within the chunk occur concurrently (e.g. *take it off*), whereas 'non-contiguous' chunks are those which have skipped items in the chunk (e.g. *take * off* where the asterisk could denote *it* or *the jacket*). More work has been carried out on contiguous chunks (e.g. Biber, 2006; Cortes, 2004; Hyland, 2008a,b), as these are more readily identifiable and the data is more easily restricted. The non-contiguous category includes Renouf and Sinclair's (1991) 'collocational frameworks'; these consist of two or more collocates or words which in some form 'belong together' in a framework (e.g. *a * of', be * to, many * of*). Such frameworks comprise a 'phraseological skeleton' (Philip, 2008: 97) as the chunks require an intervening 'content' word before they are complete instantiations (e.g. *a range of, be able to, many types of*). Often, the 'gap' in non-contiguous chunks is abstracted; that is, it is described in terms of a part-of-speech label or a semantic possibility; thus, *is a kind of, is a sort of, is a type of* can be grouped together as *is a* + N + *of* (cf. the 'grammatical patterns' of Hunston and Francis, 2000). A consideration of semantic possibilities beyond contiguous lexical chunks led Hunston (2008: 271) to develop her work in 'sequences of meaning elements' or 'semantic sequences'. An example semantic sequence is:

theory/argument + *arises from* + *the observation* + *that*-clause
(Hunston, 2008: 279)

This sequence comprises a combination of optional lexical items plus a lexical form followed by a lexical form and finally a grammatical category.

A less strictly defined, non-contiguous chunk is the concgram, defined by Cheng et al. (2009: 236) as a set of words that 'co-occur regardless of constituency variation (e.g. as AB or A*B), positional variation (e.g. AB or BA), or both'. While collocational frameworks and semantic sequences involve human intervention in the form of sorting concordance lines to find frequent frameworks, concgrams can be generated fully automatically using software such as ConcGram (Greaves, 2009) or the concgram feature within WordSmith Tools (Scott, 2010). A drawback of such searches is that extracting all possible contiguous and noncontiguous chunks of any length, and with any possible permutation of constituency or positional variation, results in an extremely large quantity of data; this then needs to be filtered in some way in order to group patterns and thus make sense of the data. While in this study non-contiguous chunks are not employed, they are referred to as a possible extension for further research. Ultimately, limitations have to be placed on the type and quantity of lexical chunks found and analyzed, in order to extract a reasonable quantity of data in relation to the research question posed.

This section has considered characteristics which are widely used in studies of lexical chunks; these are bound together with particular methods of identification and 3.3.2 considers some of these methods.

3.3.2 Methods of identifying chunks

This section looks briefly at experimental methods and use of syntactic or semantic features, before turning to a fuller consideration of intuition and frequency counts as the two main methods employed in the research literature for identifying lexical chunks.

Experiments

Some methods for identifying chunks can only be used under experimental conditions, thereby excluding naturally-occurring data. Examples include measuring brainwave activity as participants read chunks on a screen (e.g. Tremblay, 2009); recording pauses between keystrokes as participants input data (e.g. Wiktorsson, 2000, using ScriptLog software); and measuring eye fixations (e.g. Underwood et al., 2004, using eye-tracking software). However, such experimental methods might not 'identify' chunks but, instead, test whether a pre-determined list of chunks are processed faster than non-chunks. While useful for establishing the existence and range of chunks (e.g. Erman, 2007; Jiang and Nekrasova, 2007; Wiktorsson, 2000), studies using experimental methods are less useful in research on naturally-occurring language produced by learners.

Presence of particular syntactic/semantic features

A list of lexical chunks is sometimes decided on through use of a prescribed set of syntactic and/or semantic features. For instance a researcher may class as chunks only those strings which fulfil a pre-determined degree of lexicogrammatical fixedness or non-compositionality. In practice, this method is usually bundled with others; for example Simpson (2004: 42) extracted contiguous n-grams (or computationally-derived chunks) from a spoken academic corpus using frequency of occurrence, then narrowed the list to a subset of chunks fulfilling her requirement of 'structural and idiomatic coherence'. This led Simpson to include *in terms of* and *I think that* and to exclude *and in fact you* and *I know what I*, as she deemed the former to be syntactically and semantically coherent or 'whole' chunks, while the latter were found to be frequently-occurring yet were not coherent wholes, according to her intuition. Although this method has the potential limitation of reliance on an individual's intuition, this can be countered by recruiting additional language users to gauge the resulting chunks; moreover, chunks can be filtered to include only those viewed as of pedagogic value. However, this method is less suitable for large-scale analysis of corpora as each type has to be checked.

Human intuition

Categorizing chunks through their semantic/syntactic features cannot always be carried out through a predetermined set of features as it is difficult to specify all possibilities in advance. Instead, some recourse is likely to be made to human intuition, or the hard-to-define human sense that the words in a chunk together constitute a complete unit. Native speakers have a huge capacity to store and access lexical chunks as illustrated by Moon's (1998a) findings that in a 200 million word corpus there might be only one occurrence of some fixed idioms, yet these can be readily identified by NSs. However, using human intuition as a method of identification presents problems since there is wide variation in the extent to which items are recognized as wholes (as found by Foster, 2001; Leedham, 2006; Nesselhauf, 2003). For example, in a study of NNS writing, Foster (2001) asked seven NS linguists to intuitively identify chunks, accepting as formulaic those sequences identified by at least five of the seven NSs. Informants found this a difficult and tiring task, reporting 'difficulty in knowing where exactly to mark boundaries of some lexical chunks… as one could overlap or even envelop another' (p.84). In an earlier study, (Leedham, 2006) I found a correlation between the time taken by NS informants to classify chunks and the number of chunks they identified in the transcribed language of two NNSs; moreover, increasing the time taken resulted in fewer 'missed chunks' where participants agreed (retrospectively) that an item was a chunk. Nesselhauf (2003: 228) also attempted to use NS judgments, but found that not only were these variable but there were also differences between participants' acceptability judgments and their own usage of chunks.

Using intuition relies on agreement between raters as to what constitutes a chunk since the same notions of semantic unity or syntactic completeness are not held by everyone; moreover NSs may be unable to determine chunks which are valid for a NNS (as Foster, 2001, found). Intuition is hard to quantify, and is often based on a sense that a group of words 'sound right' together. One person's intuition clearly differs greatly from another's due to uncertainty as to which words 'sound right' and exactly where the chunk begins and ends. This difference could be due to their linguistic primings (Hoey, 2005) and as such differs according to each person's communicative experiences. However, providing specific

guidelines as to the boundaries of chunks would reduce the freedom of an individual's intuition and impose the researcher's view. Despite these inherent difficulties in the intuitive identification of chunks, many studies rely on intuition at some level, whether for the initial extraction of chunks or to refine a computationally-produced list of chunks (e.g. Baigent, 2005; Erman, 2007; Leedham, 2006; Li and Schmitt, 2009; Nesselhauf, 2003; Schmitt et al., 2004; Simpson, 2004; Wiktorsson, 2003; Wray and Namba, 2003).

Frequency counts

Counts of the numbers of times two or more words appear together can be easily produced by software packages for corpus analysis (e.g. WordSmith Tools, Scott, 2010; WMatrix, Rayson, 2008b) and render this method the most viable for extracting chunks from large quantities of textual data. Chunks can be readily extracted following the setting of parameters such as the length of the chunk, the minimum frequency of occurrence, and the dispersion of texts it must be found in. However, parameter-setting involves a degree of subjectivity (as these are essentially arbitrary judgments) and is usually carried out according to the pragmatic measure of how many chunks are generated under a particular group of settings. Too few chunks would result in insufficient data to analyze, too many may overwhelm the researcher and make it hard to assess the results (Schmitt et al., 2004).

Although n-grams derived on the basis of frequency alone often cut across structural boundaries, some analysts have suggested that they are still psycholinguistically valid units (e.g. Biber, Conrad and Cortes, 2004; Nesi and Basturkmen, 2006). For example Nesi and Basturkmen (2006: 286) argue that 'lexical bundles that occur with very high frequency across a range of texts are likely to be stored in memory as unanalysed chunks'. This seems a far from clear-cut conclusion to draw, since frequency alone does not ensure psycholinguistic validity. For instance it could be the case that two chunks, each of which is semantically 'whole' and (presumably) possesses psycholinguistic validity, frequently occur together with variations within each chunk (e.g. *on the other hand* and *the* + NP). A frequency-based search could then produce the commonly-occurring words from the end of one chunk and the beginning of the next, yet this hybrid chunk may not intuitively appear to be 'valid' (e.g. *the other hand the*). A simple cut-off figure of counts per million words may

lead to all instances being extracted as chunks, when in fact the frequency of the 4-gram (or four-word chunk) *on the other hand* would be higher than the frequency of 4-grams containing only part of this chunk. A way of verifying the holistic validity of chunks retrieved through frequency is to apply a statistical measure of collocation such as the Mutual Information (MI) test[16]. This test measures the extent to which the observed frequency of co-occurrence differs from what might be (statistically) expected, that is, the strength of association between words. MI works less well with very low frequencies, and in these cases the t-score is a more reliable measure since this takes raw frequencies of occurrence into account. The t-score is a measure of the confidence with which an association between words can be asserted, given the frequency of each word in the corpus. In this study, however, I have not used tests for the strength of association between words in a chunk. Instead, I have extracted chunks using keyword analysis (choosing the log likelihood test), and setting minimum counts for frequency of occurrence and measures of dispersion across texts, writers, and disciplines (see discussion of these measures in 4.3). This use of frequency counts is in line with many previous studies of lexical chunks in written language (e.g. Biber et al., 1999; Hyland, 2008a,b; Schmitt et al., 2004).

A study by Schmitt et al. (2004: 124) researches the 'unspoken assumption' of equivalence between chunks extracted through frequency of occurrence and those derived through intuition. They compiled a list of candidate recurrent chunks of varying frequency, length and function; these were a mix of chunks which seemed whole and complete (e.g. *as a matter of fact*) and those which seemed less likely to be stored holistically (e.g. *in the number of*). The resulting 25 chunks were then woven into a story which was recorded and played individually to 75 people (a mix of NSs and NNSs). Participants were asked to repeat lengthy sections of the story, the intention being to overload working memory and force them to reconstruct stretches of text. The results of this study showed that most, but importantly not all, of the chunks which had been identified as frequent in corpus studies were reliably produced by the NS participants. This finding led Schmitt et al. to conclude that while many frequently-occurring chunks clearly do align with holistic chunks (as would be expected since

16 See discussion of MI and t-score tests on the Collins Wordbank site here: http://wordbanks.harpercollins.co.uk/Docs/Help/statistics.html.

frequently-occurring language is produced by people), the two categories are not identical. However, the story used in this study is likely to have influenced the results since inconsistencies in the genre and the rather contrived language may have caused participants to falter in their recounts. The forced storyline, consists of the narrator picking up a hitchhiker who proceeds to read out an advert and a story from 'Cosmopolitan' magazine. However, the rather unnatural use of chunks found in academic writing as well as those found mainly in speech results in the inclusion of unnatural features such as the lack of contraction at the start of (1) below and the change in style in (2), both extracts from the hitchhiker's dialogue (the frequently-occurring chunks are italicized):

(1) 'It *is one of the most* relaxing things in the world, isn't it?'
(2) 'Would you pay that? Look. This one, *as shown in Figure* 1 opposite'.

The artificiality of this story is caused by the inclusion of all 25 chunks in a single text, and the mixing of chunks found in speech with those more commonly occurring in writing.

Human intuition and computationally-derived frequency counts are probably the two most widely-used means of extracting chunks in current research studies. While there is likely to be some overlap in the chunks produced through each method, there may be a considerable number of chunks found through frequency counts and not found through intuitive means. Schmitt et al. (2004) thus call for the use of both corpus and psycholinguistic approaches to complement each other in seeking an explanation of language use.

3.4 The view of lexical chunks adopted in this study

Thus far, Chapter 3 has outlined the many characteristics used for identifying lexical chunks, discussed linguistic theories surrounding the nature of chunks, and examined methods for identifying and classifying chunks. This section considers the resulting plethora of terminology for referring to chunks, detailing the rationale behind the terms used in this study.

From the previous discussion on different features of lexical chunks, it is apparent that they can be long idioms or short exclamations, fixed expressions or slots and fillers. Given the diversity of characteristics which can be employed or rejected in notions of chunks, and the ensuing range of methods for identification, it is perhaps 'little wonder that different researchers have … seen different things, resulting in a variety of terminology to express different perspectives' (Schmitt, 2004: 3). As Wray (2002) points out, a wide range of names have been given to co-occurring linguistic elements, including 'lexical chunks' (Lewis, 2000; 2002); 'lexical phrases' (Nattinger and DeCarrico, 1992); 'fixed expressions and idioms' (FEIs) (Moon, 1998a) 'collocational frameworks' (Renouf and Sinclair, 1991); 'formulaic sequences' (Wray, 2002, 2008); 'lexical bundles' (Biber et al., 1999); 'n-grams' (Milton, 1999); 'lexical bundles' (Biber et al., 1999); 'concgrams' (Cheng et al., 2006) and 'semantic sequences' (Hunston, 2008). Some terms incorporate strict frequency thresholds and distributions (e.g. lexical bundles) while others entail idiomaticity (e.g. formulaic sequences). In addition to the multitude of different terms from a range of theorists, several writers have also changed their usage of terminology over time, for example Moon (1998b) uses 'phrasal lexemes', then 'fixed expressions and idioms' in her book of this name (1998a). Often one term incorporates several others within it; for example, Conklin and Schmitt (2008: 86) discuss formulaic sequences as 'a broad cover term, including all of the various types of multi-word units and collocations, including, but not limited to, idioms, lexical bundles, and lexical phrases'. Wray (2002: 9) provides a more comprehensive list of terms used to refer to lexical chunks.

Given the array of terms and combinations of features these entail, it is essential for each study to detail the features entailed within the employed terminology, and to detail the means of extraction. In this study, the overarching term used for all linguistic items which co-occur in some way is 'lexical chunks'. This label is deliberately all-encompassing; covering such items as patterns in language with intermediate variability, frequently-occurring sequences of words and collocations or words which 'predict one another, in the sense that where we find one, we can expect to find the other' (Durrant, 2008: 5). Within lexical chunks, the key distinction made in this study is between chunks which are semantically unified or intuitively

'whole' and those which are not. This distinction was chosen as it is a point of division between theorists who believe chunks are intuitively-determined, psychologically 'whole' linguistic items, and those who regard frequency of occurrence as of greater importance. The method for identification is therefore part of the approach, since semantic unity cannot be reliably automated whereas frequency of occurrence relies solely on computer software for extraction. In this study, I employ two of the commonly-used terms within each of these schools of thought; these are outlined below.

Intuitively whole chunks are termed 'formulaic sequences' or 'formulaic language' following Wray (2002) and also Schmitt (2004). These are broad-ranging terms which include all prefabricated language regardless of length or composition. Thus, for Wray:

> [A formulaic sequence is] a sequence, continuous or discontinuous, of words or other meaning elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar (Wray, 2002: 9).

Wray (2008) elaborates on this earlier definition, distinguishing between the formulaic sequence which has the potential for being prefabricated, and the 'morpheme equivalent unit' (MEU) which has been proven to be a pre-constructed chunk of language. The MEU is thus:

> a word or word string, whether incomplete or including gaps for inserted variable items, that is processed like a morpheme, that is, without recourse to any form-meaning matching of any sub-parts it may have (Wray, 2008: 12).

In this study, formulaic sequences are viewed according to Wray's 2002 definition as holistically-produced chunks; that is, they are linguistic units which are stored and processed as wholes in the mental lexicon rather than built up analytically. No further speculation is made as to whether such sequences are morpheme-equivalent strings since it is not within the capacity of this study to research the psychological validity of the MEU. Formulaic sequences are viewed as semantically 'complete' or 'whole' units of language, such as *on the other hand, in other words, as a result* (all examples that follow are from this study).

I use the term ‘n-gram’ (and thus ‘3-gram’, ‘4-gram’, and so on) for contiguous chunks which are defined solely by frequency of occurrence. N-gram is similar in meaning to Scott’s ‘clusters’ (2008) and Biber’s ‘lexical bundles’ (1999), though each study provides slightly different parameters to constrain the resulting data. Further parameters for these frequency-based terms are specified in 4.3.4. Whereas a formulaic sequence has a particular function and context attached to it, an n-gram is limited to a frequent string of wordforms which may or may not possess a coherent function (e.g. *at the end of, that there is a, on the other hand*). Although frequency of occurrence cannot be assumed to be equivalent to internalization in the mental lexicon, there is likely to be overlap between the two sets (e.g. *on the other hand* and *at the same time* are both formulaic sequences and n-grams in the corpora since they are semantically ‘whole’ and frequently-used).

Lexical priming theory argues that the collective primings of a community of language users result in a shared inventory of language deemed to be ‘natural’, implying that frequency-based corpus searches have some validity. Omitted from frequency-based searches are those formulaic sequences which are not used frequently enough across a wide enough range of individual’s language productions to satisfy the parameters set. This could be because the corpus is simply not large enough for particular sequences to meet the designated lower frequency threshold. N-grams which are not semantically ‘whole’, however, can become functionally ‘complete’ with the addition of lexis at one or both ends of the chunk. For example, Figure 3.1 shows the ten occurrences of the n-gram *that there is a* from a corpus of 280,000 words (the Chinese corpus used in this study). On its own, this 4-gram does not seem ‘complete’.

1	Phillips curve. Friedman	argued	that there is a	specific short-run Phillips
2	& Rhodes 2005: 195). Maslow	argues	that there is a	'psychological growth'
3	corresponding LSD value, it is	implied	that there is a	significant difference
4	shows a sigmoid shape. This	indicates	that there is a	decrease in rate of
5	seem irrelevant' or when they	insist	that there is a	contract, they 'ignore
6	and x are shown in Fig. 6. It	represents	that there is a	null point C at the centre
7	However, the recent figures	shown	that there is a	drop in numbers at two
8	Although past performance	shows	that there is a	decreasing trend of the
9	people. Nevertheless, it also	shows	that there is a	strong correlation between
10	Secretary Alastair Darling had	stated	that there is a	strong argument to pause

Figure 3.1 that there is a

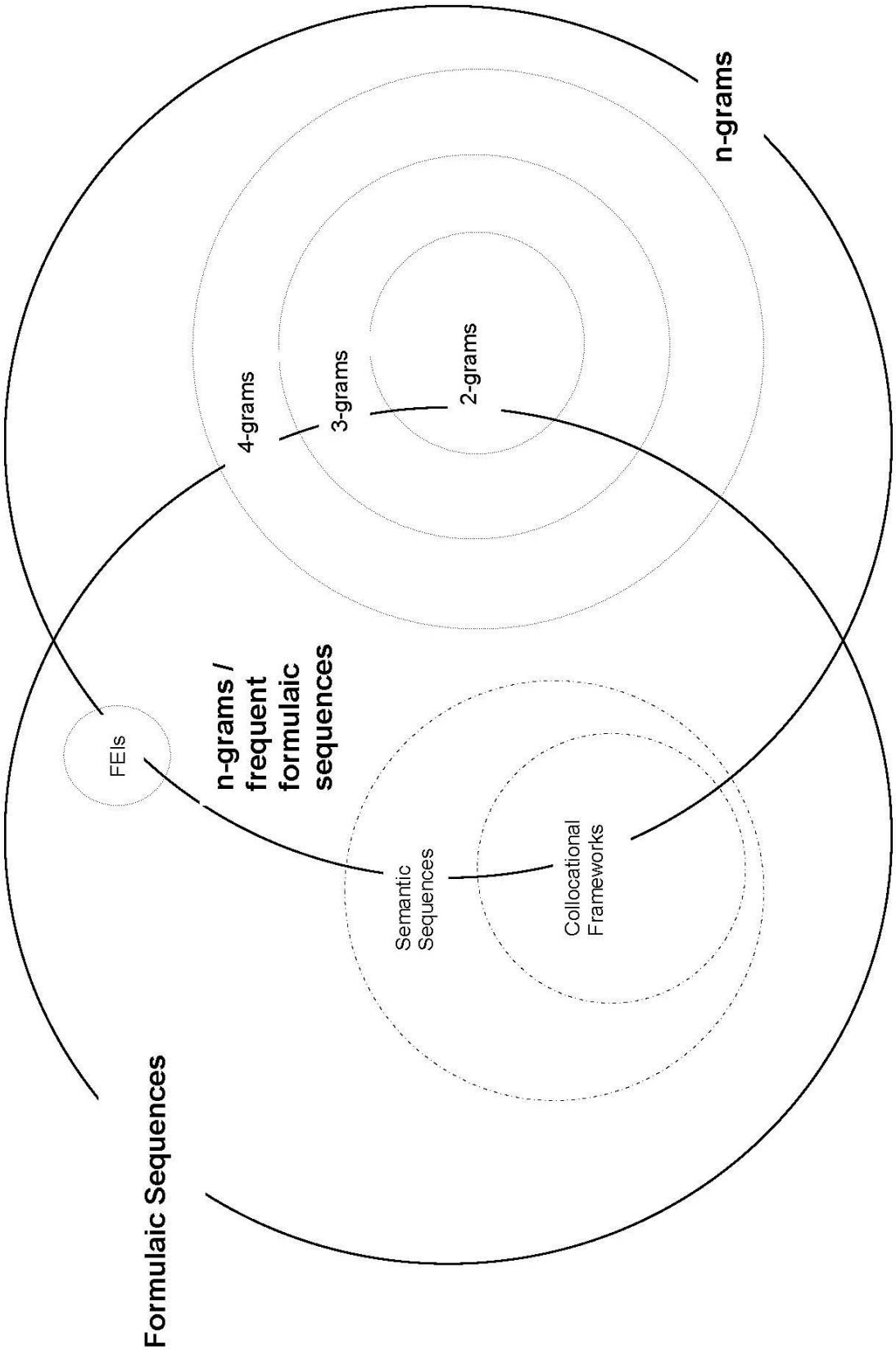
In each example, the wordform preceding the chunk is a verb, however none of these occur frequently enough to satisfy the threshold of 3 occurrences of an n-gram. If the threshold were lowered to two occurrences, then the 5-gram *shows that there is a* would be extracted. However, this 5-gram does not feel intuitively 'whole'; that is, it does not convey a full meaning on its own. Instead, meaning hinges upon the preceding and following words. Alternatively, longer chunks from Figure 3.1 could form a single semantic sequence:

NP + VP *that there is a* + NP-proposition

Semantic sequences (Hunston, 2008) are potentially very useful means of pedagogically explaining chunks. However, this thesis is primarily concerned with describing a variety of academic writing through the extraction of n-grams, rather than exploring the psycholinguistic reality behind retrieved frequency-based chunks. Some mention is made of semantic sequences, but this phenomenon is not explored in depth.

So far this section has explained and exemplified my use of the general term 'lexical chunks' and the overlapping areas of n-grams, formulaic sequences, and semantic sequences. Figure 3.2 illustrates how these labels fit within other commonly-used terms in the literature. The left-hand circle represents formulaic sequences and the right-hand one shows n-grams. Within the overlap of the two circles are examples of chunks which are both frequently-occurring and semantically-whole, such as frequent connectors (e.g. *on the other hand, in the long run, at the same time*). In the left-hand circle but overlapping slightly with the right-hand one are Fixed Expressions and Idioms (FEIs) (Moon, 1998a) (e.g. *kith and kin*); these

Figure 3.2 Overview of lexical chunks



can be frequent or infrequent, but are all contained within the circle of semantically-unified formulaic sequences. Also within the left-hand circle but overlapping with n-grams are semantic sequences (Hunston, 2008), shown here within a dotted circle to indicate the abstracted and thereby permeable nature of these chunks. Semantic sequences are incomplete structures, requiring lexis to instantiate each example and subsume the category of collocational frameworks (Renouf and Sinclair, 1991). For example, semantic sequences include *a* + noun-classifier + *of* + noun-category, instantiated as *a kind of* experiment; and the subsumed collocational framework is *a * of*, giving rise to *a kind of, a form of*.

In the right-hand circle of Figure 3.2 but overlapping with formulaic language are categories of frequently-found n-grams as these may or may not be semantically whole units; here, 2-grams are shown as contained within 3-grams, and so on (e.g. *on the* is found within *on the other* which is found within *on the other hand*). Solely in the n-gram circle are those chunks which are frequently-found but which are not semantically whole (e.g. *the other hand the*).

This study uses naturally-occurring data in quantities large enough to exclude the possibility of relying solely on human intuition. N-grams are therefore identified on the basis of frequency, using WordSmith Tools software. The semantic unity, or otherwise, of the resulting chunks is discussed further within functional classification taxonomy given in 3.5.2.

3.5 Classification of chunks

Once identified and extracted, a useful next step is to categorize lexical chunks in order to group chunks with different forms into a more manageable number of categories and to compare these categories across datasets. This categorization is reported in Chapter 6 in a comparison of year 1/2 and year 3 of each student corpus. Most researchers carry out both structural and functional classification of n-grams (e.g. Biber and Barbieri, 2007; Cortes, 2004, 2006; Hyland, 2008a,b; Hyland and Tse, 2005; Stubbs and Barth, 2003). Early categorizations blurred the distinctions between form and function, giving taxonomies containing a mix of these. For example Becker's (1975: 71) six-item list includes 'metamessages' and 'sentence builders'; and Nattinger and DeCarrico (1992: 1) describe

'conventionalized form/function composites'. More recent research has used classifications based on either the grammatical structure of chunks and/or a functional grouping and this section considers each of these.

3.5.1 Structural classification

One of Sinclair's (1991) reasons behind his 'clean-text policy' (or preference for using unannotated corpora) is that, 'although linguists leap effortlessly to abstractions like 'word' … and beyond, they do not all leap in the same way, and they do not devise precise rules for the abstracting' (1991: 21-22). Although Sinclair's note of caution should be borne in mind, structural labels can still be usefully employed both for ease of reference and to enable comparison with other studies, as long as the terms of reference are clear and agreed-upon. Thus, if chunks are classified on the basis of the word class of the first part of speech item (the usual method adopted in structural classification), then it is possible to syntactically group chunks of any kind, irrespective of whether the chunk is holistic (i.e. a formulaic sequence) or simply frequent (i.e. an n-gram).

Most categorizations have been carried out on contiguous chunks (e.g. Biber, 2004; Cortes, 2004, 2006; Hyland, 2008a,b). In the case of non-contiguous chunks, structural categories are often used to restrict the 'gap' to a grammatical part of speech (e.g. *the* N *of the* where the noun could include *nature, aim, role, impact*); the 'gap' can be further restricted to a count noun. One widely-used system is Biber et al.'s (1999: 997) categorization of chunks (employed by, for example, Chen, 2009; Cortes, 2004; Cortes, 2006; Hyland, 2005b, 2008a,b). This system was also chosen for the current study as it is extremely thorough, identifying 14 categories and giving dozens of examples with clear explanations of the differences between the categories.

Biber et al. list categories according to chunks found mainly in conversation and those found mainly in academic writing. Conversation is described in the LGSWE as more verb-based, with chunks comprising the 'building blocks for verbal and clausal structural units' (p.992). The low frequency of nouns in conversation is attributed to the lower density of information

(Biber et al., 1999: 66); clauses are short, and have a strong focus on action. In contrast, academic prose is more noun-based, with chunks forming 'building blocks for extended noun phrases or prepositional phrases' (p.992). A focus on nouns is more prevalent since academic writing is concerned with a high density transmission of information. Table 3.1 shows the categories in the LGSWE which are relevant to my study, omitting those 'more widely used in conversation' (p.997) except for category 3 (8% of 4-grams in conversation, 2% of 4-grams in academic prose).

Broad category	**Structural pattern**	**Examples from Biber et al., 1999**
NP-based	(1) NP + *of*-phrase fragment	*the end of the, the form of the, the importance of the*
	(2) NP + other post-modifier fragment	*the way in which, the relationship between the, an important role in*
	(3) pronoun/NP (+aux) + *be*	*it was in the*
PP-based	(4) PP + embedded *of*-phrase fragment	*as a result of, by the presence of, in view of the*
	(5) other PP fragment	*at the same time, in the present study, on the other hand*
VP-based	(6) anticipatory *it* + VP/AdjP (+complement clause)	*it is possible to, it is necessary to, it can be seen*
	(7) passive verb + PP fragment	*is shown in table, referred to as the, be used as a*
	(8) *be* + NP/AdjP	*is one of the, is the same as, may be due to*
	(9) (NP+) (verb +) *that*-clause fragment	*should be noted that, studies have shown that, that there is a*
	(10) (V/Adj+) *to*-clause fragment	*are likely to be, has been shown to, to do with the*
Other	(11) other expressions	*as well as the, may or may not, than that of the*

Table 3.1 Structural classification of chunks (adapted from Biber et al, 1999: 997)
Key: Adj P = adjective phrase, NP = noun phrase, P*P*= prepositional phrase, VP = verb phrase

The left-hand column of Table 3.1 is headed 'broad category' and follows Chen and Baker's (2010: 34) grouping of the structural patterns as 'noun phrase-based', 'preposition phrase-based', or 'verb phrase-based' as these are a useful way to compare the smaller categories

(though note that I keep to Biber et al.'s structural patterns whereas Chen and Baker merge the two patterns in the 'NP-based' category and have an additional VP-based category). Hyland and Tse (2005) also use Biber et al.'s classification, commenting that academic writing uses a high number of 'NP + post-modifier fragments' (e.g. *the number of, the relationship between the*), 'preposition + *of* fragments' (*in terms of, on the basis of*) and anticipatory *it* fragments *it was found that, it should be noted that*).

3.5.2 Functional classification

Categorizing the functions of chunks has played a significant role in the study of phraseology in both learner corpora and NS studies (e.g. Biber and Barbieri, 2007; Chen, 2009; Cortes, 2004, 2006; Hyland, 2008a,b; Hyland and Tse, 2005; Stubbs and Barth, 2003). As Spöttl and McCarthy assert:

> All approaches to the study of formulaic language stress the importance of their functional aspect, that is, the fact that certain language sequences have conventionalized meanings which are used in certain predictable situations (2008: 208).

Academic writing can be viewed as one such set of predictable situations. For example, Cortes (2006) discusses chunks in the context of acquiring disciplinarity, arguing that frequent use of chunks:

> …seems to signal competent language use within a register to the point that learning conventions of register use may in part consist of learning how to use certain fixed phrases (2006: 398).

An example of this is use of the chunk *through the prism of* in an undergraduate Sociology assignment[17] to signal the methodological view (or 'lens') through which phenomena are observed. Use of this sequence could signal membership of the whole community of academic writers or perhaps one narrowed to social scientists. Conversely, avoidance of all such preferred ways of saying things would distance a student writer from the academy and lessen the sense of belonging to the group.

[17] An L1 English interviewee studying Sociology highlighted this as a chunk which was new to him.

In the context of a functional approach, one of the most widely-used theories has been Halliday's (1994, also Halliday and Matthiessen, 2004) Systemic Functional Linguistics (SFL). This theory of language differs from the majority of those discussed in 3.2 in its greater emphasis on paradigmatic choices (that is, items which can be substituted) over syntagmatic ones (items which occur within a construction or pattern). SFL places importance on how paradigmatic options allow the speaker to 'represent the world in a particular way, to construe a particular relationship with the hearer, and to weight information in particular ways' (Hunston, 2006: 65). In contrast, a phraseological approach 'brings attention back to the syntagm, in the sense of the sequence or pattern' (Hunston, 2006: 65). Where SFL and phraseological approaches share common ground, however, is in placing context to the fore, with language users seen as making lexical choices that can be explained by the social context (Tucker, 2005, provides further discussion on SFL and phraseological choices).

Studies taking a functional approach usually analyze chunks using a three-way taxonomy, along the lines of the metafunctions identified in Halliday's Systemic Functional Grammar, though not using his labelling of 'interpersonal', 'ideational' and 'textual' metafunctions (Biber et al., 2003; Hyland, 2008a,b; Pecorari, 2009). Nattinger and DeCarrico (1992) adopt similar tripartite categories in their consideration of the functions of chunks, using the headings: 'social interactions', 'necessary topics' and 'discourse devices'. Biber, Conrad and Cortes (2004) identify 'stance expressions', 'discourse organizers' and 'referential expressions'. Cortes (2004) follows Biber et.al. (1999) and considers 'referential bundles', 'text organizers', 'stance bundles' and 'interactional bundles', the last two of which she groups together as 'interpersonal functions'. Hyland (2008a,b) draws on Halliday (1994) and Biber et al. (2004) for his taxonomy of 'research-oriented', 'text-oriented' and 'participant-oriented' chunks. However, whereas Biber et al. (1999) investigate frequent chunks in a wide range of both spoken and written language, Hyland narrows the focus to academic writing. This study employs Hyland's functional categories as these have been modified for academic writing and also enables my corpora of student assignments to be compared with Hyland's

professional corpora. The three divisions of Hyland's functional taxonomy are described below.

Participant-oriented chunks provide a 'structure for interpreting a following proposition' (Hyland, 2008b: 18). 'Engagement' features are reader-focused, serving to draw the reader into the discussion through direct address. In contrast, 'stance' features are writer-focused and reveal the writer's views and the extent of their commitment (see Table 3.2).

Meta function	Specific function	Examples from Hyland, 2008a,b
Participant-oriented	Engagement: address readers directly,	*as can be seen, it should be noted that*
	Stance: convey the writer's attitudes and evaluations	*are likely to be, may be due to, it is possible that*

Table 3.2 Participant-oriented metafunctions (based on Hyland, 2008a,b)

Research-oriented chunks function, in Hyland's (2008b: 13) words, to 'help writers to structure their activities and experiences of the real world' (see Table 3.3).

Meta function	Specific function	Examples from Hyland, 2008a,b
Research-oriented	Description: describe elements of the research	*the structure of the, the size of the, the surface of the*
	Location: indicate time/place	*at the beginning of, at the same time, in the present study*
	Procedure: show how the research was carried out	*the use of the, the role of the, the purpose of the, the operation of the*
	Quantification: indicate the quantity of elements in the research	*the magnitude of the, a wide range of, one of the most*
	Topic: relate to the field of research	*in the Hong Kong, the currency board system*

Table 3.3 Research-oriented metafunctions (based on Hyland, 2008a,b)

In the same way as Halliday's ideational metafunction, research-oriented chunks concern the research itself, describing concrete or abstract parts of it ('description'), where and when it takes place ('location'), how it is carried out ('procedure'), and the number of concrete or abstract parts ('quantification'). Also included is the subcategory of 'topic' chunks related to the research field; while these are extremely frequent as locally-repeated chunks in a single assignment, they are infrequent within a corpus (assuming there is a reasonable range of different topics).

Text-oriented chunks are similar to Halliday's textual metafunction and Biber's 'text organizers' in that they are connected with the overall organization of the text (see Table 3.4). These chunks may situate the research (or knowledge in general) in context ('framing'), show how aspects of the research relate to each other ('resultative'), organize the discourse ('structuring'), or signal whether part of the discourse is similar or different from its neighbours ('transition'). Chunks in the latter subcategory are very likely to be taught in both EAP and general English classes and to appear in textbooks as they often form coherent wholes such as *in the long run* and *on the other hand.*

Meta function	**Specific function**	**Examples from Hyland, 2008a,b**
Text-oriented	Framing signals: situate arguments by specifying limiting conditions	*in the case of, with respect to the, on the basis of, in the presence of, with the exception of*
	Resultative signals: mark inferential or causative relations between elements	*as a result of, it was found that, these results suggest that*
	Structuring signals: text-reflexive markers which organize stretches of discourse or direct the reader elsewhere in text	*as shown in figure, in the present study, in the next section*
	Transition signals: establish additive or contrastive links between elements	*on the other hand, in addition to the, in contrast to the*

Table 3.4 Text-oriented metafunctions (based on Hyland, 2008a,b)

Hyland (2008b) argues that some of these functional categories are 'strongly connected' to structural patterns. For example he found that 'NP + *of*' structures were common in the

research-oriented category; ‘prepositional phrase patterns’ within the text-oriented functions; and ‘anticipatory *it*’ patterns within the participant-oriented functions. Mappings between structural and functional categories are discussed further in 6.6.

Difficulties of a monofunctional classification

The process of assigning chunks to functional categories is to some extent subjective in that one linguist’s ‘participant-oriented’ chunk may be another’s ‘text-oriented’ one, and it is therefore important for the researcher to present a comprehensive list of the categories, complete with examples from the data. Much of the difficulty of functional categorization, however, is due to attempts to assign all instances (or tokens) of a type to a single metafunction. Researchers following Hyland’s classification have described their difficulties in adopting this monofunctional classification (e.g. Pecorari, 2009; Oakey, 2009). For example, Pecorari (2009) discusses her procedure for classifying n-grams in a corpus of Biology research articles. After some deliberation, she decided that at least 50% of the tokens of any n-gram had to be related to a particular function for the n-gram to be assigned to that functional category. She found that ‘a small number of bundles’ were employed for two or three functions with ‘more or less equal frequency’ and these were not allocated to any category (p.94). However, Hyland does not discuss any difficulties in categorizing the chunks in his studies (Oakey, 2009, also makes this point). This lack of discussion makes Hyland’s work difficult to replicate, as the basis for categorization is not always clear.

Instead of assuming a single functionality, it could be argued that many chunks are in fact multifunctional and can be classed within two or even three of the metafunctional groupings (cf. Moon’s, 1998a: 241, discussion of what she terms ‘cross-functioning’ whereby ‘a speaker/writer uses FEIs in functions other than their canonical ones, thereby foregrounding or thematizing the selection’). A single chunk could have an interpersonal component (determined by, for example, the presence or absence of a modal) and/or a textual component (e.g. the repetition of an earlier sequence), and/or a research-oriented one (relating to the topic of the prose). For example, the chunk *can be seen as* could be categorized as both interpersonal since it introduces a comment from the writer, and textual, as it is often used to refer to tables and diagrams in the data. Assuming chunks have a

single function is probably over-simplistic. While there may be a dominant function for a type, closer examination of the instances of the chunk may reveal a secondary function, and subsuming all tokens under one metafunction can distort the findings.

Note that the examples in Table 3.3 (from Hyland's data) reveal that the students are from Hong Kong and suggest that business or finance is a frequent topic (one of the four disciplines in Hyland's corpus is in fact Business Studies). Although all of Hyland's (2008a,b) postgraduate student data is from Hong Kong Chinese students, he does not comment on how their L1 could have influenced the choice of n-grams. The focus of his research is on academic level (masters, PhD or professional academic) and discipline (Electrical Engineering, Business Studies, Applied Linguistics and Biology), with language taken as a constant across these. The issue of writers having English as an L1 or L2 is barely mentioned. Difficulties in functional classification are further discussed in 4.3.4.

This section has discussed two taxonomies for categorizing n-grams: one of these is based on their structural form and follows Biber et al.'s (1999) comprehensively explained and exemplified system. The second taxonomy is taken from Hyland (2008a,b) and categorizes n-grams according to their function; however, this is less clear-cut as Hyland provides fewer examples, and he assumes that each chunk has a single dominant function. The discussion in 3.5.2 has critiqued this assumption and explored some of the difficulties experienced in attempting to classify in this monofunctional way.

3.6 Chapter summary

This chapter began by discussing lexical chunks as an increasingly common way of comparing NNS writing, and as a fundamental aspect of language (e.g. studies by Chen, 2009; Chen and Baker, 2010; Gilquin and Paquot, 2007; Granger, 1998; Paquot, 2010; Hyland, 2008a,b; Lee and Chen, 2009; Li and Schmitt, 2009; Thompson, 2009; Wiktorsson, 2003). The chapter then examined how a range of linguistic theories tackle the area of phraseology, beginning with a description of Miller's (1956) information processing theory and how its notion of storing information in larger and larger chunks paved the way for later

theories which applied the concept of chunks to language. Hopper's (1987, 1998) emergent grammar has similarly been adopted in later theories, with Sinclair (1991), Hunston and Francis (2000), and Hoey (2005), all building on the idea of fluid rather than fixed grammars. Hoey's theory of lexical priming was used in the chapter to draw together aspects of other theories, and is seen as (broadly) compatible with Wray's (2002) needs-only analysis, Hunston and Francis' (2000) pattern grammar and Sinclair's (1991) idiom principle. Key to the application of lexical priming to student writing is the acceptance that a discourse community's shared language patterns are derived from the shared primings of individuals.

For this study, the broad term of 'lexical chunks' is divided into computationally-derived 'n-grams' (which are frequent but perhaps not semantically 'whole') and the psycholinguistic category of 'formulaic sequences' (which, while coherent chunks, might not be frequent and are difficult to discern from data). As this study concerns large amounts of data, n-grams are extracted based on pre-selected frequency parameters (i.e. number of occurrences and dispersion across texts), and no claims are made for the psycholinguistic reality of the ensuing chunks. Finally, the chapter considered the structural and functional categorization of the resulting n-grams, following taxonomies by Biber et al. (1999) and Hyland (2008a,b) respectively.

As discussed in this chapter, comparisons of the phraseology of large quantities of texts have been facilitated by corpus explorations, and Chapter 4 turns to the data for this study and the corpus linguistic procedures used to analyze them.

CHAPTER 4 DATA AND RESEARCH METHODS IN THE STUDY

4.1 Introduction

Chapter 3 discussed lexical chunks as the starting point for the analysis in this study and this chapter considers the data and research methods used. This introductory section provides a rationale for the primary focus on Corpus Linguistics, contrasting this with qualitative methods for analyzing student writing.

Student writing, whether short argumentative essays or lengthier undergraduate and postgraduate assignments or dissertations, has often been studied through qualitative means. Qualitative analyses include studies of particular lexical items (e.g. Myers, 2001, researched students' use of *in my opinion*); investigation into how features of text interrelate (e.g. Lea and Street, 2006, studied multimodal resources used by students writing within different genres), and research into how far students draw on previous academic writing experience (e.g. Whitley, 2007, conducted an interview study on the academic development of a sample of international students). This type of analysis is often carried out on small numbers of students and texts, and may include richly-detailed descriptions of student attitudes to writing in order to understand the process of text production (e.g. the ethnographic research of Lillis, 2001). Datasets tend to be small due to the intensive nature of the investigation and, while providing insights into features of individuals' writing and into their attitudes and previous learning, may therefore be limited in terms of the generalizations which can be made.

In contrast to these small-scale studies, using the methodology of Corpus Linguistics extends the possibilities of linguistic research beyond that which 'a single individual [can] experience and remember' (Sinclair, 1991: 1), enabling large quantities of data to be analyzed in a relatively short time. Rather than concentrating on a few individuals, corpus analysis enables the investigation of groups of texts from a range of individuals (e.g. from a

sample of Chinese undergraduate students) and thus 'moves away from individual preferences to focus on community practices, dematerializing texts and approaching them as a package of specific linguistic features employed by a group of users' (Hyland, 2009: 110).

In addition to the quantity of data which can be comfortably analyzed, a corpus study can reveal unexpected patterns. In corpus-driven studies, researcher bias is reduced since areas of interest emerge from the texts without recourse to preconceptions as to linguistic forms or functions (i.e. the data 'drives' the research), rather than resulting from searches for pre-existing categories of lexis or grammar (as is the case with corpus-based enquiry) (Tognini-Bonelli, 2001). Corpus linguistic identification tools thus do not simply speed up the process of pattern identification, they enable the indication of language patterns that the human reader might skim over and either fail to notice or dismiss as unremarkable. Yet comparing how the commonplace varies across texts is exactly what is required in determining how writing varies across student groups, over time, and between disciplines. Whereas non-corpus methods of text analysis can only be carried out on small numbers of texts, due to the time required, and risk excluding less obvious phraseological patterns in the writing, corpus analysis, due to the organization of large quantities of data, brings these patterns to the forefront. As Sinclair (1991: 100) states, 'the language looks rather different when you look at a lot of it at once'.

While recognizing that it is impossible to forego all prior ideas, this study takes a corpus-driven approach to allow patterns to emerge from the data. Using a keyword[18] procedure reduces researcher bias as search terms are not *selected* for the data but *emerge* from the keyword procedure and are then followed up through further searches (e.g. concordance lines, dispersion plots, collocate lists) and through qualitative means (selection, categorization of keywords and reading of the surrounding text). Keyword analysis also assists in reducing the influence of previous findings from the research literature as the automated nature of extracting keywords lowers any researcher bias (though bias cannot be eliminated since intuition is always needed to make sense of keywords). A corpus-driven

18 Here, 'keyword' is restricted to words occurring statistically more frequently in one corpus than in another, not 'keyword' as in 'keyword in context' or KWIC concordance.

approach is thus highly compatible with a focus on lexical chunks as it allows lexico-grammatical patterns to emerge from the data.

In this chapter I first describe the textual data employed in the study and the process of preparing the data for analysis. The chapter then outlines the corpus linguistic procedures used to sift the data, namely n-gram extraction, keyword and key n-gram lists; and the automated means used in analyzing the extracted data, for example concordance lines, dispersion plots and collocate lists. Finally, I discuss the additional procedures of detailed multimodal analysis of pairs of assignments and extraction of relevant points from lecturer interviews (from the BAWE project) employed in Chapter 7.

4.2 The data

This section outlines the data used in the study, that is, the BAWE corpus data and the additionally-collected data, and describes the process of selecting from these datasets to produce the final corpora used in the study. Compiling a corpus, whether the data is primarily extracted from an existing corpus (as is the case here), or constructed anew, involves many subjective decisions and it is important to be transparent as to what has been decided (cf. Hunston, 2002: 123).

As well as the corpus data, various other datasets such as class observations, interviews, and questionnaire responses were collected during the period of this study, and these have informed and shaped my views on Chinese students' assignment-writing in UK universities. The datasets include my observations of French language lessons in a UK secondary school; and observations of English language lessons in Beijing primary schools, secondary schools, universities and a private language school. Semi-structured face-to-face and email interviews with UK and Chinese undergraduate students (n = 12) were carried out which explored attitudes towards UK and Chinese university writing. Attendance at Chinese language classes for 18 months gave me a measure of insight into differences between constructing written text in Chinese and in English. An online questionnaire was completed by 200 Chinese and English UK university students and, although the responses to

questions on assignment-writing are not discussed in this thesis, the questionnaire served as a means of encouraging students to send me their assignments for inclusion in my research, and provided some insights into the range of genres in undergraduate assignments and the difficulties experienced. Informal discussions with questionnaire and interview respondents, and other Chinese and British students at varying levels of academic study have also informed my views on assignment-writing in UK HE.

The main source of data remains the BAWE corpus, and the next section describes this dataset.

4.2.1 The BAWE corpus

The BAWE Corpus was collected at the Universities of Oxford Brookes, Reading, Warwick and Coventry within the framework of the ESRC-funded project, 'An investigation of genres of assessed writing in British Higher Education' (Nesi et al., 2005). The finished corpus comprises around 6.5 million words within approximately 2,900 student assignments from over 30 disciplines and four levels of study (three undergraduate years and one masters year). All contributed texts were processed through Turnitin plagiarism detection software to ensure, as far as possible, that they comprised students' own writing and were not copied from textbooks or internet sources. Constraints were set on the number of words per assignment (minimum of 500 and maximum of 10,000), thus allowing the inclusion of short coding exercises within Computing at the lower end and undergraduate dissertations at the upper end. Further constraints were on the number of texts which could be submitted by an individual (up to five assignments from one level of study and up to ten from a single discipline).

The corpus has an 'inclusive approach to academic writing' (Sharpling, 2010: 192) in that writing from both L1 and L2 English students is included, as long as it fulfils the criterion of proficiency (as discussed in 2.2). A further area of inclusiveness which is noted by Sharpling is the range of universities included from the more traditional to the former polytechnic

institutions (created from 1992), and the consequent range of vocational programmes of study (e.g. Publishing; Hospitality, Leisure and Tourism Management).

The primary data of the BAWE corpus is supplemented by contextual data on both contributor (e.g. date of birth, gender) and on assignment (e.g. grade, module title); this information is stored in the document header of each student text. In line with ethical practice and in order to protect the anonymity of participants, all identifying data is removed. Following submission, individual assignments were first converted from Word to text documents and then saved under a four-digit number followed by a letter. The number uniquely denotes the individual contributor and the letter denotes the individual text; thus student 0254 gave text 0254a, 0254b, and so on. The next stage in the BAWE compilation process was the mark-up of the files. This mark-up follows the guidelines of the Text Encoding Initiative (TEI) (Sperberg-McQueen and Burnard, 2004), resulting in TEI-conformant XML files. The mark-up itself was carried out semi-automatically in the case of the document title, table of contents, sections and subsections, lists, block quotes, formulae, abstract or summary, bibliography/references and appendices; and automatically in the case of sentence, paragraph, highlighting, and section numbering (Ebeling and Heuboeck, 2007).

4.2.2 Additional assignments

In total, 245 assignments (576,153 words) were submitted to the BAWE corpus by Chinese students. Although this is a substantial number from one L1 group, once assignments are allocated to year groups and disciplines the number of individuals and wordcounts per group is much reduced. To increase the numbers for the present study I attempted to collect supplementary assignments from Chinese students in a range of UK HEIs. However, this proved to be both difficult and time-consuming. The BAWE project was a £450,000 research project, which took place over a three-year period, with a personnel of nine (four full-time equivalent researchers and five academics), and funding to pay students £3 per submitted assignment. In comparison, I did not have the contacts across a range of universities, and could not provide contributors with a financial incentive. However, as I was not limited to a

few participating universities I could appeal for submissions across a wider spread of UK Higher Education institutions.

I used varied means of attracting contributors. These included posting messages on discussion fora aimed at either Chinese students or at English language teachers, circulating emails to current and former teaching colleagues with Chinese students, advertising in Open University publications for Associate Lecturers to ask students they taught in other institutions, and attending and handing out flyers at relevant conferences. Among the more successful of these approaches were a personal contact at Lancaster University, emails to the Chinese Scholars and Students Association website, and notices posted on the social networking site Facebook. The latter site was successful as its popularity was greatly increasing at the time of collection particularly among university students (Ellison et al., 2007). Within Facebook I posted requests to relevant groups, set up a group, and regularly posted (free) notices and (paid for) adverts targeted at students in universities with high numbers of Chinese students (as detailed on the Higher Education Statistics Agency website [HESA, 2010]).

For all electronic collection methods, interested students were asked to click on a link and were taken to an online page with information on my project[19] and could then contribute assignments through email. Contributors were asked to supply contextual information, in line with the data requested for the BAWE corpus, and were assured of anonymity. Following consultation with the OU Research Ethics Committee, students were also asked to include a statement in their email giving explicit consent for their assignment to be used for research.

As well as collecting assignments from individual students or via individual tutors I researched existing corpora of Chinese student writing. However, these were all deemed unsuitable for addition to the corpora for this study due to the short length of essays (as low as 200 words), limited nature of the writing (argumentative and narrative essays only) and, most importantly, the fact that the writing was conducted in a non-UK environment (see

19 Students could also choose to complete a ten-minute questionnaire on assignment-writing, for which they were offered the small enticement of entry into a £20 prize draw. The questionnaire responses are not discussed in this study.

discussion of learner corpora in 2.3). Other corpora were compiled from specially-written texts, for example the Chinese portion of ICLE collected at Portsmouth University consisting of 500-word 'argumentative' essays (Papp, 2009). An additional 50 assignments from Portsmouth University were collected from 50 year 3 undergraduate Chinese students of Business but these were limited to 'practice essays' written on controversial topics (Papp, 2008, personal communication). A further specialized corpus of 75 Engineering texts from Liverpool John Moores University was offered to me but not included in my corpus as the students were Malaysian Chinese and all assignments were on the same topic within Artificial Intelligence, so would skew the dataset.

The only texts from existing collections that were included were four assignments from a Lancaster/London School of Economics writing project (Whitley, 2007) and eight from a PhD research project at Lancaster (Kinzley, 2011). In all, a total of 133 additional assignments were collected from 18 UK universities (95 texts from L1 Chinese students and 38 from L1 English students[20]). The additional assignments were from all three undergraduate levels and masters level and covered a range of disciplines. However, most of the texts by Chinese students were at masters level (see Table 4.1).

Chinese	**texts**	**tokens**
from BAWE (masters and undergraduate)	245	576,153
additional (masters)	73	239,847
additional (undergraduate)	22	50,524
total	**340**	**866,524**

Table 4.1 Number of tokens and texts per student corpus (before refining)

The same process of anonymization and deletion of all references, appendices, diagrams, tables, charts and pictures was carried out as for BAWE corpus preparation. Contextual

20 Although the BAWE corpus contained sufficient texts from L1 English students, for the sake of parity across institutions I did not exclude these from the search for additional assignments. However, in the final corpus, only five L1 English additional texts were included.

information on each student and each text was added to a combined Excel spreadsheet of the extracted BAWE texts and the additionally-collected texts.

4.2.3 Refining the corpora

This section describes the process of gradually limiting the BAWE corpus and additionally-collected data from all undergraduate and masters level assignments to undergraduate assignments from the same disciplines and written by L1 English and Chinese students only. Decisions as to which texts to include in a corpus are also pragmatic, and in this study reflect the contextual information collected from participants and the number of texts available within different categories in the BAWE corpus. Contextual information in file headers and an Excel database lists each student's self-determined L1 (as with IELTS data, information on country of birth/upbringing was not requested from participants).

In corpus analysis, a tension exists between the selection of texts from as homogeneous a group of writers as possible, and the compilation of a sufficiently large body of data for analysis to be meaningful. An example of a homogeneous group would be L1 Mandarin speakers whose entire secondary education took place in one province of the PRC and who have no prior study abroad. A maximally similar group would reduce the number of variables and allow more confident assertions to be made as to the characteristics of the writing of a particular group, whereas a more heterogeneous corpus is likely to give access to a far larger quantity of data, and enable greater generalization to be made. Since the number of texts by Chinese students would be too restricted if the parameters were tightly-defined, the study thus took a pragmatic approach and included a broader range of texts from 'Chinese students' of different nationalities and dialects, which fulfil the selection parameters. This data was then narrowed to texts from undergraduate students who have spent the majority of their secondary education in their home country (whether the PRC, Hong Kong, or other).

A pilot corpus was first compiled which consisted of a single text per student contributor. This followed a statistician's advice[21] that most statistical tests are based on each element

21 From the Open University Statistics Advisory Service.

containing data from a single individual and thus more conclusive results could be drawn from following this model. However, limiting the corpus in this way involved many minor decisions as to which texts to choose, based on length, discipline, year group, and thus introduced a greater degree of subjectivity. While the benefit of lessening the effect of individual variables is apparent (and is one of the advantages of the ICLE learner corpus), restricting the data severely curtails the corpus size. For example, Chen's (2009) work on Chinese students' texts form BAWE adopted a one-text-per-student model, including undergraduate and masters level texts (without differentiation) and thus limited the data from Chinese students to just 53 texts (146,872 words). However, after some trialling, this model was not followed due to the limitation of the data and the merging of levels, and the corpus was enlarged to include most Chinese student contributions, subject to the restrictions outlined below.

The first limitation to the data outlined in Table 4.1 was thus to exclude the masters level assignments by Chinese students (n=155, wordcount = 501,255) as these are substantially different to undergraduate work. This difference is in part because Chinese masters students are likely to come to the UK purely to undertake their course rather than taking a postgraduate course after a UK-based undergraduate one, making their assignments less viable as samples of a 'fourth' year of study. Moreover, the available masters assignments from Chinese students are from a different array of disciplines, with greater weighting towards 'soft' disciplines (such as Education and Applied Linguistics[22]) and a lower proportion of writing from 'hard' disciplines (such as Engineering and Computing). Although many masters level students undertake their degree as an end in itself, for others it might be regarded as training for an academic career in a way that undergraduate writing is not. For these reasons, the masters assignments are viewed as qualitatively different to the undergraduate ones and were not included in this study. However, the texts were prepared in the same way as the undergraduate assignments and could be used in a follow-on study.

22 However, in contrast to this, figures provided by the British Council, 2010, on the most popular postgraduate subjects for Chinese students in 2008/9 include Business, Finance, Economics, Management, Electronic and Electrical Engineering, but *not* Applied Linguistics or Education.

In addition to the removal of postgraduate assignments, those texts from students with four or more years of secondary education outside the country of their birth and early upbringing were also excluded in order to reduce the number of students with a substantially different educational background (including, for example, 'British Born Chinese'). A four-year cut-off was chosen as it seemed likely that some Chinese students interpreted the question of study abroad as including their previous three years of undergraduate study in the UK (see Appendix B for example of the contextual data form completed by students).

From the remaining texts, it seemed that combining levels of undergraduate study might be a fruitful way of reducing the categories of comparison (from years 1, 2, and 3) and thus simplifying the data. I decided to combine years 1 and 2 of undergraduate study, and to compare these with year 3. The rationale for this is fourfold: in part it is practical as the wordcount is roughly the same in the combined texts from Chinese students in years 1 and 2 as from Chinese students in year 3. In part, the decision was a response to factors surrounding the nature of UK study as sometimes 'year 3' assignments are from fourth year students who had an industry placement in year 3. Additionally, a small number of the Chinese contributors are termed '2 + 2' students, that is, they studied for two years in a Chinese university and then joined year 2 in the UK. Thus, their work is classed as 'year 2' but is from their first year of UK university study. Finally, year 3 writing could be construed as qualitatively different from that produced in earlier years as assignments tend to be longer and include additional genres such as research reports. It is likely that tutors expect a higher quality of writing in year 3, that is, there may be a qualitatively greater jump from year 2 to year 3 than from year 1 to 2 (BAWE interview data).

Table 4.2 illustrates the reduction in wordcounts from the original BAWE corpus by the application of successive criteria (stages 2, 3, 4), then the supplement of the additionally-collected texts (stage 5) to the final corpora in stage 6.

1	All BAWE texts from undergraduate years 1, 2, 3 plus masters year (6.5 million words)				
2	All undergraduate texts (4.7 million words)				
3	Texts from writers with no more than 4 years of secondary education abroad (4.4 million words)				
4	English (3.1 million words)				Chinese (230k)
5	English (same disciplines as Chinese) (1.3 million words)		Chinese from BAWE (230k)	+ Extra Chi (50k)	
6	'Eng12' (876,894)	'Eng3' (458,782)	'Chi12' (140,341)	'Chi3' (139,354)	

Table 4.2 Progressive refinement of datasets (not to scale)
'abroad' here denotes outside country of birth and early upbringing

The total wordcounts in the resulting corpora of Chinese undergraduate texts (the focus of this study) and in English undergraduate texts (comprising the reference corpus) are given in stage 6 of Table 4.2. From the final total of 146 texts (280,000 words) by Chinese students used in this study, 25 (46,000 words) are from Mandarin speakers, 40 (61,000 words) from Cantonese speakers, and the remaining 81 (173,000 words) are from unspecified L1 'Chinese' students. Students in the latter category perhaps believed that in the UK 'Chinese' is equated with 'Mandarin' and so failed to specify further, while Cantonese-speakers perhaps wished to assert a separate L1 identity. It is also the case that many PRC and Singaporean students are bi-dialectal, using Mandarin as the language of education in school and a different dialect of Chinese in the home, and may be unclear as to which of these constitutes their 'first language'. Of the Cantonese-speakers in the study, many are likely to be from the Guangdong region (one of the most populous provinces of the PRC and one in which Cantonese is prevalent) as well as from Hong Kong. As noted in Chapter 1, PRC students outnumber Hong Kong students in UK Higher Education Institutions by over 4:1 (47,035 PRC students compared to 9,600 Hong Kong students) (HESA, 2010). The majority of students in this study are therefore likely to be L1 Mandarin speakers from the

PRC, with a smaller number of L1 Cantonese-speakers from southern provinces of the PRC and Hong Kong.

4.2.4 The final data

This section describes the final corpora; these are referred to as 'Eng12', 'Eng3', 'Chi12' and 'Chi3', denoting the L1 of students and the year groups of undergraduate study. As described in 4.2.3, the criteria for inclusion of texts in the Chinese corpora were the student's first language and years of study abroad; and for individual texts, the year of study, single authorship, grade and length of text. The English dataset had the same criteria, with the extra requirement that disciplines not represented in the Chinese corpora were removed (e.g. Linguistics, Classics, History). Other variables of students and texts were not taken into account in the compilation of the corpora but are given here in order to assess the similarity of the student groups. Details of the corpora overall are given in Table 4.3 and the division by year groups is shown in Table 4.4.

	Chi123	**Eng123**
Number of tokens	279,695	1,335,676
Number of texts	146	611
Number of students	45	70

Table 4.3 Number of tokens and texts per student corpus

	Chi12	**Chi3**[23]	**Eng12**	**Eng3**
Number of tokens	140,341	139,354	876,894	458,782
Number of texts	89	57	436	175
Number of students	30	20	45	34

Table 4.4 Number of tokens and texts per year group corpus

23 The Chinese year group corpora were almost the same size and the decision was taken to omit two texts from students in Chi3 who contributed large numbers of assignments, thereby rendering Chi12 and Chi3 each around 140,000 words.

There are far more English texts in the study since there is a far larger pool of English students to draw on, resulting in many more contributions to BAWE. For each of Chi123 and Eng123, the majority of texts are from students in the 18 – 27 age group, with a small number of texts contributed by students aged from 28 to 48 (6 texts for Chi123 and 19 texts for Eng123). As well as there being fewer mature students, it is perhaps likely that students in this age range are busier and thus less keen to contribute assignments. In terms of grades, the two student groups are similar with roughly equal numbers of distinctions (see earlier discussion in 2.4 on the nature of 'proficient' writing)[24]. The contextual feature where the two student groups vary the most is in the gender proportions of the corpora. In Chi123, texts from women students outnumber those from men by a ratio of 5:2 (103 texts from female students, 42 texts from male students, one where the gender is not given). However, in Eng123 the proportions are almost equal (307 texts from female students, 304 texts from male students). It is not clear why there is this disparity, as this gender difference is not reflected in the number of female and male Chinese people studying in the UK as a whole (according to the British Council, 2010b, the gender proportion of students from the PRC enrolled on undergraduate programmes in the UK in 2008/9 are 53% female and 47% male). I can only speculate that the gender difference in contributing texts may be due to a difference in willingness to participate in a research project.

Wordcounts in disciplines

The wordcounts per discipline in the four corpora are shown in Table 4.5 as percentages of each corpus. This illustrates some comparability (e.g. Biology comprise between 10 and 15% of each corpus, Mathematics is never greater than 2% of a corpus) but also the domination of some disciplines in particular corpora (in particular, Cybernetics is 17% of Chi3 but just 3% of Eng12 and Eng3).

24 A small number (n=4) of the additionally-collected Chinese assignments were later found to be slightly under the 60% required but were left in the corpus.

Discipline/%	Chi12	Chi3	Eng12	Eng3
Agriculture	0	7	12	9
Biology	14	10	15	10
Business	13	7	5	9
Computing	1	6	7	8
Cybernetics	0	17	3	3
Economics	13	16	4	4
Engineering	18	8	15	15
Food Science	12	10	6	4
HLTM	11	5	5	5
Law	8	14	16	13
Mathematics	0	1	2	2
Sociology	10	0	11	17
Totals	100%	99%	98%	98%

Table 4.5 Wordcounts per discipline as a percentage of each corpus (NB Totals do not all add up to 100% due to rounding)

The division of 'soft-hard' and 'pure-applied' discussed in 2.5 gives a useful paradigm for this study as it is helpful to see the most common groups of disciplines for Chinese students. Figure 4.1 illustrates the dispersion of the 12 disciplines along Biglan (1973) and Kolb's (1981) 'hard-soft' and 'pure-applied' dimensions, following their classifications as far as possible both in selecting the quadrant and in the precise location within each quadrant. While these classifications are now somewhat dated, they remain the most comprehensive set of discipline categorizations and are a useful starting point. Where particular disciplines were not present in either Biglan or Kolb's original classification, the nearest approximation was taken. For example, 'Food Science' was not in the original matrices; this was judged to be closest to Biglan's 'Dairy Science' and was accordingly placed in the same location within the 'hard applied' quadrant. Some discrepancies between the classifications remain; for instance Kolb places Law in the 'soft applied' quadrant, while Becher (1989) suggests this belongs within the 'soft pure' category. In the case of Law, I have followed Kolb's decision, as this seems to better reflect the practical nature of many of the Law modules. Figure 4.1 is

intended to provide an approximate location of the disciplines only, since it is beyond the boundaries of this study to precisely locate where a discipline as constructed within a particular university department is situated. It should also be borne in mind that students themselves can select optional modules from a particular end of the spectrum, perhaps rendering their degree more 'applied' than would be considered usual.

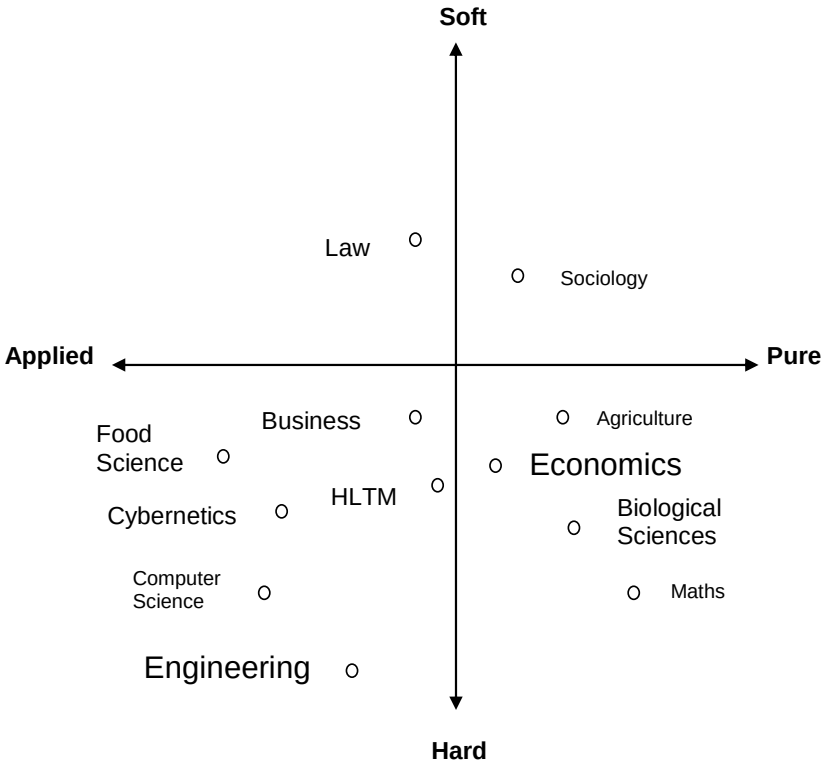


Figure 4.1 Disciplines in the study, arranged on 'hard-soft', 'pure-applied' dimensions (based on Biglan, 1973; Kolb, 1981). The font sizes in the figure relate (approximately) to the wordcounts of each discipline in Chi123.

While remaining a somewhat rough outline of where a broad discipline area may be placed on the 'soft-hard' and 'pure-applied' dimensions, Figure 4.1 provides a visual overview of the range of disciplines studied by the Chinese participants in this study, indicating that Chinese students are far more likely to study disciplines towards the 'hard' end of the 'soft-hard' dimension. Within this, the 'hard-applied' quadrant contains the most disciplines and also the largest wordcount overall (indicated approximately by the font size of the named disciplines, with precise figures given in Table 4.5). The assignments within BAWE and the additionally-collected ones are from a wide range of disciplines and the concentration of Chinese students' texts in the lower quadrants in Figure 4.1 indicates a preference among this group

for ‘hard’ and ‘applied’ disciplines. This preference accords with data on the British Council website (2010) stating that Chinese students in UK universities most commonly take undergraduate courses in Business/Management, Finance/Accounting, Engineering and Computing Science.

It is likely that two of the factors which influence the concentration of Chinese students’ discipline choices within the ‘hard applied’ quadrant are funding and language level. First, since Chinese students’ parents usually fund the high cost of overseas university education (Gieve and Clark, 2005), students wish to show a clear return on this investment by opting for applied disciplines such as Engineering or Business which lead to a specific career rather than more generalist Arts and Humanities disciplines. In some cases, it might be the parents who choose their child’s discipline area and thus future career choice. Second, ‘softer’, more discursive disciplines in which language is at the forefront often demand a higher English language score than those in which language plays a lesser role, leading Chinese students to favour ‘hard’ disciplines over ‘soft’ ones. For example, in the International English Language Testing System (IELTS), a linguistic proficiency test commonly taken by international students, Reading University requires a score of 6.5 for most undergraduate courses, but 7.0 for entry into Arts and Humanities, Education or Law courses. Similarly, at Warwick, the Faculty of Science asks for IELTS 6.0 but the Faculty of Social Science requires 7.0. It is likely that recounting events and presenting data are highly valued in ‘hard’ disciplines and writing in continuous prose is more important in ‘soft’ disciplines. This point is made succinctly by Neumann et al.:

> a skill with deploying facts and figures counts for more than elegance of writing style: many students survive scientifically-based courses with very little need for skills in prose exposition (2002: 412).

Similarly, Warburton (2006: 7) has argued that ‘in humanities subjects … students are judged on their essays. If you can’t write good essays, especially under exam conditions, then you will never succeed in these areas’. An ability to write lengthy prose seems, then, to be valued more in the ‘soft’ disciplines than in the ‘hard’ disciplines.

Genres and genre families

In addition to considering the proportions of different disciplines in the corpora, it is also worth considering the proportions of 'genres' of writing. Unlike the uniform nature of texts in a learner corpus (similar-length, argumentative essays), undergraduate texts may have a argumentative, narrative, reflective, recounting, or other primary function. Swales (1990: 58) defines a genre as 'a class of communicative events, the members of which share some set of communicative purposes', which suggests there are particular conventions or rules connected with a communicative event. Within Applied Linguistics the study of genre has had a great deal of attention (see, for example, Johns, 2002 for discussion on genre in educational contexts); despite this, there is a lack of agreement as to precisely what genre entails, leading Hyland (2003: 213) to aptly describe it as a 'slippery concept'. In the BAWE project, genres were identified and classified in Heuboeck et al. (2008) in an iterative process of reading assignments, assigning them to potential genre categories, and checking for consistency. Each genre was also classified within one of 13 different 'genre families' which cut across disciplines and disciplinary groupings for the purpose of permitting 'ready comparisons across disciplines' (Gardner, 2008: 20). For example the genre family of 'explanation' should 'demonstrate understanding of the object of study; and the ability to describe and/or assess its significance' and includes such genres as 'business review' (found in Business and in HLTM), 'species/breed/overview' (from Biology), 'product development overview' (e.g. in Engineering) and 'system/process overview' (Computing) (Heuboeck et al., 2008: 48; a full list of the 13 genre families with example genres is given in Appendix C). The process of slotting assignments into genre categories is, however, a subjective one since it relies on manual reading and grouping of texts, and as such differs from one individual to another. It is difficult to move away from preconceived notions of what a particular genre entails. Moreover, the extent to which a single genre can cut across disciplines is unclear since the same notional genre may differ from one discipline to another (this point is also made by Bruce, 2010, in his discussion of essays in Sociology and English, and reviewed in 2.5).

I initially attempted to group the additionally-collected assignments, using in Heuboeck et al.'s (2008) 13 genre families. However, this proved difficult and led me to believe that these categorizations should be used with caution since the genre and genre family categories cut across previously-used genre categories. For example the commonly-used genre label of 'reflective writing' is not used in Heuboeck et al.'s categorization; instead, most writing which might be subsumed under this heading is named 'reflective recount' and appears in the genre family of 'narrative recounts'. However, reflective writing consisting of a 'reflective letter to a friend' is also found in the 'empathy' family. Reflective writing is an important category in UK HE (as evidenced by the increasing research within this area reported at conferences such as Writing Development in Higher Education) and accounts for some of the discrepancy in the use of the first person pronouns across assignments. The re-naming of reflective writing to 'reflective recounts' within the genre family of 'narrative recounts' proved problematic in my later analysis. Although Heuboeck et al. allow for compound genres (e.g. 'essay + narrative recount') within the same assignment, I found additional cases where reflective writing is included as a small section of a non-compound text. An additional problem for this study was that introducing yet another means of categorizing texts (along with L1, discipline, year group) had the effect of minimizing the available data in any single category of the corpora. In order to compare like with like, should, for example, L1 Chinese Economics year 3 reports only be compared with L1 English Economics year 3 reports? This is clearly not viable in a relatively small corpus and, after trialling the use of genre family categories, I decided not to prioritize genre categories in my analysis.

4.3 Corpus linguistic procedures

The corpus linguistic procedures used in the study were carried out using the most recent version of Scott's WordSmith Tools (released in 2008 and updated frequently[25]) as this provided the requisite functionality (cf. discussion of software in Wiechmann and Fuhs, 2006). The WordSmith package comprises three tools: Concord, Keywords, and Wordlist, all of which are used in this study. The corpus linguistic procedures employed included

25 The latest version of WordSmith Tools (version 5) is used throughout this study. This was released in 2008; however, as it is updated monthly the date given for the version is that used in the analysis: namely, 2010 (as suggested by Scott, 2010, help files).

describing characteristics for each dataset, lists of keywords and key n-grams (collectively referred to as 'keywords'), counting types and tokens, and extracting and categorizing lists of 4-grams. These procedures are described in turn below.

4.3.1 Describing characteristics of the texts

The first points of comparison between the corpora are in the overall statistics of mean assignment length (MAL), that is, the mean number of words per text; mean sentence length (MSL), measured by the number of words per sentence; and mean word length (MWL), measured through the number of characters per word. Together, these statistics provide a set of characteristics (or 'linguistic profile') of the datasets. Use of a set of general text characteristics has some precedent in the literature (e.g. Engber, 1995; Gardner, 2009; Grant and Ginther, 2000; Jarvis et al., 2003). For example, Grant and Ginther (2000: 130-1) consider type-token ratio and mean word length as 'indicative of sophisticated writing' since, as students 'become more proficient writers, they are more precise about using words that best express their ideas' (their findings were discussed in 2.4).

Mean Assignment Length (MAL)

The first characteristic of MAL can be calculated using Excel from the number of texts and number of tokens (as given in WordSmith). At lower levels of linguistic proficiency, text length has been used as an index of writing fluency and linked to the ratings awarded (e.g. Grant and Ginther, 2000; Jarvis et al., 2003), though Jarvis et al. (2003: 400) point out that the reason for this correlation is unclear: perhaps being a good writer encourages a student to write more, or perhaps raters 'are simply biased towards longer texts'. Since all writing in the study is at a proficient level, the MAL is calculated simply to provide an overview of the datasets.

Mean Word Length (MWL)

The MWL is calculated by WordSmith as the mean number of characters per orthographic wordform and is provided in a 'statistics' tab in the 'Wordlist' tool. The 'characters' can be alphabetic letters, numerals, punctuation marks or other symbols. All strings of characters with spaces between are included in counts of words, meaning that abbreviations and

technical terms are treated as words for the purpose of calculating the MSL and MWL, with the exception of formulae (e.g. mathematical formulae, computing code) which are replaced in the textfiles by the capitalized word 'FORMULA', and each counted as a single word.

Mean sentence length (MSL)

The MSL can be calculated differently, depending on the features chosen as sentence endpoints. For example, 'strong' punctuation such as full stops, exclamation marks and questions marks are a standard means of indicating the end of a sentence, but sentence fragments are treated differently. To verify the calculation used in WordSmith I ran 18 randomly-chosen assignments from Chi123 through both WordSmith and an alternative word counting program (Wordcount, Danielsson, 2009). The two programs have slight differences in the treatment of sentence fragments. For example in Wordcount, numbered headings such as '*1*' are counted as one-word sentences, while a sentence ending in a formula followed by another sentence is counted as just one sentence rather than two (Table 4.6).

	WordSmith Tools	**WordCount**
Total Words	25,875	25,825
Total Sentences	1,212	1,232
Mean sentence length	21.34	20.96

Table 4.6 Comparison of software calculations of MSL

The overall difference in mean words per sentence between the two programs is not statistically significant (using the log likelihood test, p=.0001), suggesting that the word counting feature in WordSmith is sufficiently accurate for the purposes of this study.

The measures described in this section are used in Chapters 5 and 6 to give an overview of broad differences between the datasets. More detailed analysis is carried out through the extraction of keywords.

4.3.2 Extracting keywords

The main procedure used to explore each dataset in the study is that of keyness, in common with a range of other studies (e.g. Baker, 2004; Culpeper, 2009; Rayson, 2008a; Scott and Tribble, 2006; Xiao and McEnery, 2005). 'Key' items are those which occur statistically more often in the small corpus than the large corpus, relative to the total number of words in each corpus, meaning that keyness is thus a 'matter of being statistically unusual relative to some norm' (Culpeper, 2009: 34). This section sets out the procedure for calculating keyness using WordSmith Tools, considering issues around the choice of reference corpora and choices surrounding keyword extraction.

Keyness in WordSmith Tools

In order to calculate keyness in WordSmith, a word list is first generated through the word list tool. Using the keyword tool, WordSmith then compares the frequency of an item in the smaller wordlist with the frequency of an item in a reference wordlist. As well as searching for individual words, the keywords function can be used to search within lists of n-grams (called 'key n-grams' in this thesis; note that 'keyword' is also used at a superordinate level to denote both keywords and key n-grams). In addition to *positive* keywords, WordSmith can provide a list of items which are *negatively* key, that is, a word or n-gram which 'occurs *less* often than would be expected by chance in comparison with the reference corpus' (Scott, 2010, original emphasis). Negative keywords are commented on where these are useful to the analysis in this study, though the main focus is on positive keywords. It is also possible to extract key keywords in WordSmith, that is, words which are key in a set number of texts in the corpus, thus ensuring that a single text is not dominating the search for keyness. In this study only keywords are used, though all instances are checked to ensure they occur within a range of texts and from several individuals.

For Scott and Tribble (2006: 78), a keyword is 'an ordinary word which happens to be key in a particular text', giving an indication of the text's 'aboutness' since dominant concepts are likely to be repeated verbatim. This definition, then, excludes paraphrased concepts and items in relations of synonymy/metonymy. Extracting key concepts using semantic domains can be achieved through WMatrix (Rayson, 2008b), though this necessarily relies on pre-

defining the categories (key semantic categories using WMatrix were extracted, though as the results confirmed findings from WordSmith keyword analysis they are not reported on here). Keyness is described by Scott and Tribble (2006: 55-6) as 'what the text "boils down to" … once we have steamed off the verbiage, the adornment, the blah blah blah'. However, this definition is a text-based notion of relevance in that it assumes that meaning is constrained within the boundaries of the text(s) and is explicitly expressed within the text. An alternative view is to regard keyness as discoursal; that is, to view words or chunks as key within the text(s) and the broader environment in which they occur. The latter view thus accounts for the process of reducing lengthy keyword lists (produced by software) to shorter, selective lists (produced by the human researcher).

Choice of reference corpus (RC)

In keyword searches, the choice of reference corpus (henceforth 'RC') is significant as this affects the number and type of keywords extracted. This section reviews some of the recent literature concerning RC in keyword searches, in order to justify the RCs used in the later analysis chapters.

First, in terms of size of RC, Berber Sardinha (2004, in Scott, 2009) claims that the larger the RC, the higher the number of keywords that will be extracted, and consequently recommends an RC of approximately five times larger than the target corpus as a good starting point. Scott's (2009: 90) trials of differently-sized reference RCs offer a refinement on the issue of size, as he suggests that where a 'mixed-bag' RC is used then a large RC is better, whereas if a more specific RC is used the size is less important. The second point of consideration from the literature is that of the scope of the RC. Rayson (2008a: 527) suggests that 'it is important that the two corpora do not overlap, or that one is not a sub-corpus of the other', in order to maintain independence and thereby increase validity. The third consideration is the issue of similarity of the target corpus and the RC. In his attempt to find a 'bad RC' (in order to discern features of an inverse, 'good' RC), Scott (2009: 90) found that even an RC which is very different in genre to the target corpus can provide 'plausible indicators of aboutness', leading him to claim robustness for the keyword procedure. Culpeper (2009: 35), however, argues for similarity between target and RC, since 'the closer

the relationship between the target corpus and the RC, the more likely the resultant keywords will reflect something specific to the target corpus'. For example, comparing a highly specific corpus such as Chi-Engineering (writing in Engineering produced by Chinese undergraduate students) with a general reference corpus (such as the BNC) would give a list of keywords used by student writers, as well as those used in Engineering. In contrast, comparing Chi-Engineering to Eng-Engineering would reduce the number of keywords connected with student writing generally and Engineering in particular, and reveal differences in the Chinese and English corpora in this discipline. Baker (2004: 349) points out that a keyword analysis which uses a similar RC to the target corpus will focus 'only on lexical differences, not lexical similarities'. For example, in his research on gay and lesbian erotic texts Baker compares the texts to each other to find differences between types of erotic literature, and also compares each text to a corpus of general English to find items which are key in both the smaller corpora.

My keyword comparisons in Chapters 5, 6, and 7 take into account the various suggestions as to the validity of an RC. Thus, in Chapter 5, I compare Chi123 with Eng123. Eng 123 is (almost) five times larger than Chi123, satisfying Berber Sardinha's (2004) criterion of size. The two student corpora contain similar texts (Culpeper, 2009) yet do not overlap, nor is one a subset of the other (Rayson, 2008a). Chapter 6 uses the same data as Chapter 5, and explores how the extracted keywords are used in years 1/2 and year 3 of each student corpus. In Chapter 7, each of three discipline subcorpora in Chi123 are compared to the equivalent in Eng123 and also with the whole of Chi123 and Eng123 (barring the particular discipline to maintain independence, Rayson, 2008a). Thus, the RCs for Chi-Biol are Eng-Biol and a new RC containing Chi123 and Eng123 texts *minus* Biology tests. Sets of key items are thus pruned and given in the following groups (cf. Baker, 2004): (i) items which are key in Chi-Biol both when the corpus is compared to Eng-Biol *and* when compared with the all-undergraduate corpus; (ii) items which are key in Eng-Biol when compared to both Chi-Biol *and* when compared to the all-undergraduate corpus; (iii) items which are not key when Chi-Biol and Eng-Biol are compared with each other, but *are* key when each is compared with all-undergraduate. This enables comparison of key items across the three disciplines

(group iii above for each discipline), in addition to consideration of key items in each discipline subcorpus (groups i and ii above).

The keyword procedure

Before extracting keywords, various parameters must be set in WordSmith, namely the minimum frequency threshold, the test chosen for statistical significance, and the *p* value (or probability value) selected for this test. The probability value gives the likelihood of the result occurring by chance; for example, a p-value of 0.01 (p=0.01) means that there is a 1 in 100 chance the result occurred by chance. For the study, a frequency threshold of six occurrences for 3 to 5-grams as, following repeated tests, this was sufficient to discount most of the very topic-specific items occurring across texts in a discipline. For keywords and key 2-grams, the threshold was raised to 20 instances in order to limit the number of key items to a manageable level. The setting of this threshold (as with later manual ones) is to an extent subjective since 'there is no popular consensus about cutoff points' (Baker, 2004: 351), and where relevant, key items falling below 20 (but above 6) are examined in Chapters 5 and 6. The log likelihood test was selected to determine keyness, following Dunning's (1993) argument that chi square and mutual information tests are less valid than the log likelihood (G^2) test where counts are low. The log likelihood measures how surprising an event is, no matter how few times it occurs within a corpus, and is thus useful for research on small corpora. A conservatively 'safe' *p* value of 0.000001 (or 1 in a million) was adopted after repeated experiments found that this gave a sufficient yet manageable number of keywords (cf. the trial and error procedure reported on in Chen, 2009). Negative keywords were also extracted (i.e. items which are significantly more frequent in Eng123 than in Chi123).

It is important to bear in mind that any cut-off point is by nature arbitrary and to an extent is determined to maintain a reasonable quantity of data for analysis, whilst also enabling comparison with other research. Since the keyword procedure effectively treats a corpus as one big text, it cannot account for internal differences so it is necessary to further limit the keywords to avoid the idiosyncrasies of individual students. As the student corpora consist of multiple texts by the same individual, and contain texts from 12 disciplines, I further refined

the keywords through the additional parameters of the range of disciplines, and the number of individual students, in order to exclude keywords localized to one or two disciplines or to a few students. Thus, keywords had to occur in a minimum of three disciplines and in the writing of at least five students, though in practice most occurred in texts from far more students. These constraints cannot be set in WordSmith, and were applied through manual analysis of the concordance lines.

Following extraction of keywords and further refining through use of concordance lines, an iterative process of categorization was carried out to assign the items to 'key categories' (Baker, 2004); this drew on the literature on student writing (e.g. 'informal items', 'connectors'). The keywords were then explored through the searches and displays available, such as concordance lines for co-text, dispersion plots for a graphic display of the occurrences of an item throughout a set of text, and collocate lists to give insight into the items a word or n-gram is commonly found with. Additional searches were made of items that did not reach the keyword list but were similar, since these might not have individually reached the specified cut-offs (cf. the semantic tagging in Rayson's, 2008b, WMatrix). For example, *bit* is a keyword in Chi123 (with Eng123 as the RC); a concordance line search revealed this commonly occurs in the n-gram *a bit of* so I also searched for *a lot of.*

The keyword procedure followed in Chapter 5 is summarized below:

1. Set parameters in Wordsmith for minimum frequency threshold (20 occurrences for keywords and 2-grams, 6 occurrences for 3 to 5-grams), statistical test (log likelihood), *p* value (.000001).
2. Obtain keywords.
3. Check concordance lines manually to eliminate keywords that do not fulfil the additional parameters of a minimum dispersion of 3 disciplines and 5 students, and from both years 1/2 and year 3.
4. Categorize the (reduced) list of keywords into key categories such as 'informal items' (e.g. *a little bit*), 'references to data' (e.g. *according to, as below*).

5. Investigate the keywords e.g. through concordance lines, dispersion plots and collocate lists.
6. Conduct additional searches for semantically similar items, and for keyword and key n-grams occurring below the threshold of 20 occurrences (but above 6).

The keywords procedure described in this section provided a corpus-driven means of uncovering differences between the student corpora. The use of a shared RC for the discipline searches in Chapter 7 provided a check that the differences are not emphasized at the expense of similarities (cf. Baker's, 2004, note of caution on this). The next section describes the methodology behind counting n-gram tokens and n-gram types. While n-gram tokens can be counted and normalized per million words to take account of differently-sized corpora, it is argued that effective comparisons between corpora are harder to achieve for n-gram types.

4.3.3 Counting n-gram tokens and n-gram types

A common initial procedure in corpus analysis is to count the number of n-gram tokens and types in a corpus and to compare these with those found in an RC or in previous studies (e.g. Chen, 2009; Cortes, 2004; Hyland, 2008a,b). This section discusses the difficulties inherent in comparisons between corpora of different sizes and which consist of texts of varying lengths. Chapter 6 reports on counting n-grams of different sizes across the corpora, though findings are tentative due to the nature of authentic corpora. I do not report on counts of types in the study as the procedure was not sufficiently robust when carried out across corpora containing texts of varying lengths. While the section on counting types below does not set out a methodological procedure employed in the analysis chapters, it is included as the discussion adds to the debate in Chapter 2 surrounding the use of same-size texts created for corpus investigation (i.e. learner corpora) and investigation of naturalistic texts (i.e. the datasets for this study).

Counting tokens

Two search parameters are usually determined in studies employing token counts: the number of instances of each type per million words (or other normalized figure) and the

number of texts the type is found in. Other potential criteria include the dispersion of types across texts from a range of individuals, or from a range of genres. Use of fixed parameters can ensure comparability of findings across differently-sized corpora and across a range of texts; however, there are no agreed-upon parameters in the literature and each research project has experimented and established their own thresholds (Table 4.7 gives search parameters in a sample of recent studies of academic writing).

Study	**Size of n-gram**	**Minimum frequency threshold**	**Minimum dispersion threshold**
Biber et al, 1999	3-grams or longer.	10 pmw	5 or more texts
Biber, Conrad and Cortes, 2004	4-grams	40 pmw	5 or more texts
Cortes, 2002, 2004	4-grams	20 pmw	5 or more texts
Nesi and Basturkmen, 2006	Spoken corpus of lectures. Disciplinary groupings are those used in BAWE.	10 x per disciplinary grouping and 50 x in whole corpus	None as disciplinary grouping requirement eliminates high use of bundles from one user.
Biber and Barbieri, 2007	4-grams	40 pmw	3, 4 or 5 texts depending on the subcorpus size. N-grams with a raw count of 3 must occur across 3 different texts.
Hyland, 2008a,b	4-grams	20pmw	10% of texts
Pecorari, 2009	4-grams to 12-grams.	5 or more times in the corpus	5 or more texts

Table 4.7 Summary of frequency and dispersion thresholds used in selected n-gram studies of written academic corpora

For example, Biber and Conrad (1999) employ the relatively low minimum frequency threshold of ten instances per million words (pmw) and a dispersion threshold of five texts; Biber, Conrad and Cortes (2004) use the more conservative minimum frequency of 40 instances (and the same dispersion threshold); whereas Hyland (2008a,b) used the frequency threshold of 20 pmw and a dispersion of 10% of texts.

In deciding on the search parameters for counting tokens, I compared two ways of establishing that tokens occur across a range of texts: counting tokens occurring across a set *number* of texts (following Biber's various studies) and counting those occurring across a fixed *percentage* of texts (following Hyland). I first tried Biber et al.'s (2004) method as this seemed one of the more conservative sets of parameters. I used the WordSmith Tools wordlist function to make an index for all undergraduate texts in the L1 Chinese corpus, and searched for 4-grams, setting the frequency to 12 as this gives a minimum of 42.9 hits pmw and matches Biber et al.'s frequency threshold:

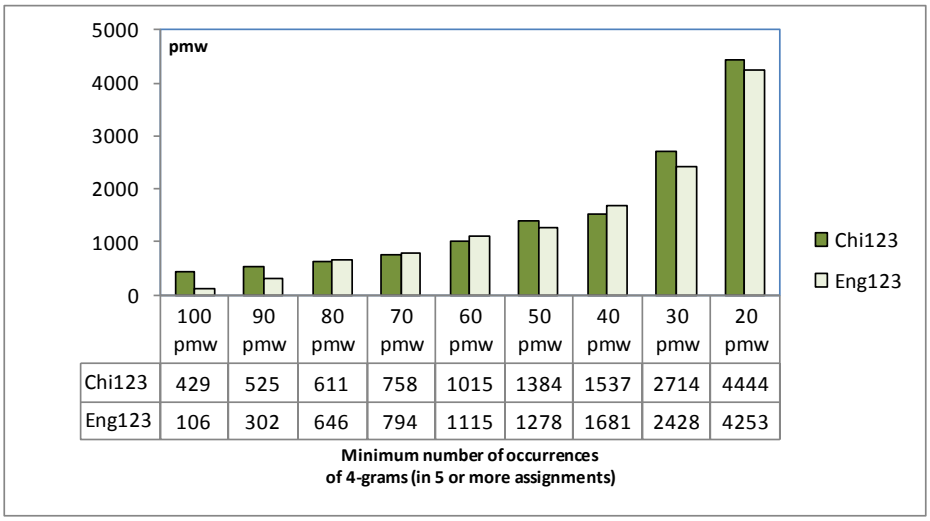
(12/279,695) x 1,000,000 = 42.9 hits pmw (and rounded down to 40pmw).

From the resulting 4-gram list I deleted all n-grams occurring in fewer than five texts, again following Biber's search parameters. This gave 22 four-gram types and 430 tokens, resulting in a normalized token count of 1,537 four-grams pmw (calculation is: [430/279695] x 1,000,000 = 1,537.38 and rounded to a whole number). For the combined English corpora, the minimum frequency was 54 hits to reach a count of 40 or more pmw:

(54/1,335,678) x 1,000,000 = 40.4 hits pmw (rounded down to 40pmw).

This gave a total of 26 four-gram types occurring over 54 times in the corpus, with each type occurring in over five texts. The total number of 4-gram tokens was 2,246 or 1,681 pmw ([2,246/1,335,678] x 1,000,000 = 1,681.54). The procedure was repeated for 5 and 6-grams; however, there are just 70 five-gram tokens reaching the 40 per million word (pmw) and five texts minimum thresholds in Eng123 and none for Chi123. No 6-grams in either corpus occur frequently enough to be included.

To explore the effect of different frequency thresholds, I calculated the normalized number of hits for each corpus, at all minimum frequency thresholds from 20 to 100 instances pmw and with each n-gram type occurring in at least five different assignments (see Figure 4.2).



	100 pmw	90 pmw	80 pmw	70 pmw	60 pmw	50 pmw	40 pmw	30 pmw	20 pmw
Chi123	429	525	611	758	1015	1384	1537	2714	4444
Eng123	106	302	646	794	1115	1278	1681	2428	4253

Figure 4.2 Comparison of normalized token counts of 4-grams (20pmw, five or more texts)

The pairs of bars show that for each student corpus as the minimum number of occurrences is reduced, the number of 4-gram tokens increases. This is as expected, since a lower threshold means more instances of each 4-gram are found. However, the difference between the two student corpora also varies: at the level of 100 occurrences pmw there are more than four times as many 4-grams extracted from Chi123 as from Eng123, whereas at 20pmw the numbers are almost the same.

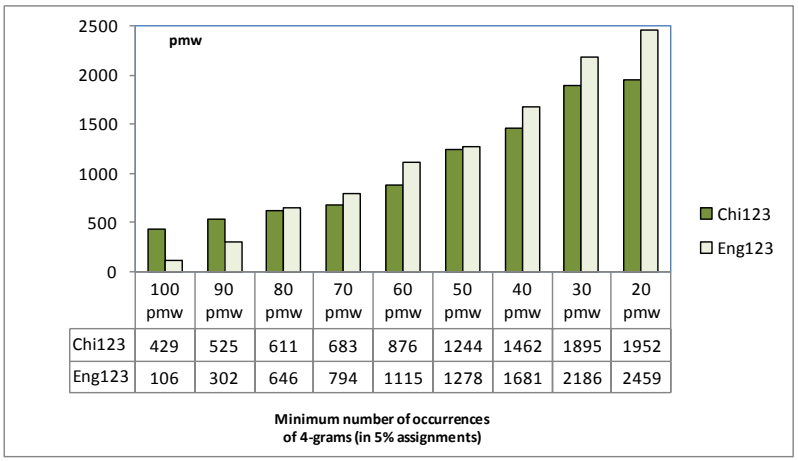
The second means of counting tokens uses the measure of dispersion across a fixed percentage of texts. Hyland (2008b) discusses the n-gram thresholds used by Biber and Barbieri (2007) and by Cortes (2004, 2006), then proposes that using n-grams occurring at least 20 times pmw and in a minimum of 10% of corpus texts provides sufficient data while maintaining a conservative frequency threshold. Searching for n-grams which occur both frequently and across a set percentage of the texts in a corpus serves to take account of the variability of text wordcounts and numbers of texts and would appear to offer a more consistent way of comparing corpora than setting a fixed number of texts. However, a search for 4-grams in my corpora using the same search criteria as Hyland resulted in very few n-grams occurring across 10% of texts. This may be due to the smaller size of these corpora in comparison with Hyland's data. As with the 20pmw criteria, the decision as to the percentage of texts in which an n-gram should be present is an arbitrary one and is based on the

quantity of data required. If a dispersion threshold of 10% of assignments is adopted, then only six 4-grams are found in the whole of Chi123 (see Table 4.8).

Rank	4-gram	Freq	Ass'ts	%
1	*on the other hand*	54	42	29
2	*at the same time*	35	24	16
3	*as a result of*	31	20	14
4	*in the case of*	27	18	12
5	*as well as the*	24	16	11
6	*the relationship between the*	20	15	10

Table 4.8 4-grams in Chi123 (20pmw, 10% texts)

An alternative threshold is to calculate the dispersion of tokens using the threshold of 5% of assignments (Figure 4.3).



	100 pmw	90 pmw	80 pmw	70 pmw	60 pmw	50 pmw	40 pmw	30 pmw	20 pmw
Chi123	429	525	611	683	876	1244	1462	1895	1952
Eng123	106	302	646	794	1115	1278	1681	2186	2459

Figure 4.3 Comparison of normalized token counts of 4-grams (20pmw, 5% texts)

A comparison of Figures 4.2 and 4.3 indicates that if a minimum dispersion rate of five or more assignments is set, then more 4-grams are found in Chi123 occurring at the rate of 20pmw; in contrast, if a parameter of 5% of assignments is set, then the number of 4-grams in Eng123 at 20pmw is higher. Additionally, for the tighter setting of 5% of assignments, the number of 4-grams is lower for both corpora.

For this study, I adopted Hyland's criterion of 20 instances pmw but relaxed the range of texts to accept n-grams occurring across 5% of texts rather than Hyland's 10% in order to give a reasonable amount of data to work with. I additionally checked that each n-gram occurred in the writing of a minimum of three different students within each corpus (whether student, year group or discipline-specific). This eliminated some n-grams within Chi123 which occurred multiple times in multiple texts, yet on closer examination turned out to be in texts written by the same one or two individuals. The number of 4-gram tokens is discussed and the most frequent fifty 4-grams are examined and categorized in the consideration of writing across the year groups (Chapter 6).

Counting types

The search parameters described above work well for token counts, as using the same criteria of N counts pmw enables normalized token counts to be compared across corpora of different sizes. Thus in Chi123 (approximately 280,000 words) a single n-gram must occur a minimum of six times to achieve a normalized rate of 20pmw but 27 times in Eng123 (approximately 1.3 million words) to achieve the same normalized rate. However, the same process cannot be applied to type counts since as a corpus increases in size, the number of different types does not increase at a uniform rate. Biber and Barbieri (2007: 269) discuss a method for the normalization of types in corpora of different sizes and containing different numbers of texts. Following experimentation across their corpora of varying genres of academic texts, they suggest that for a 50k corpus, lexical bundles should occur across a minimum of three texts; for 100k, across four texts; and for 200k, across five texts. Their searches produced similar numbers of n-gram types from each corpus and they argue that the method is thus a valid means of comparing types across corpora of different sizes.

To clarify the workings of the method I followed their procedure on randomly-chosen texts from my largest corpus, Eng12. I first compiled subcorpora of four different sizes from 50,000 words to 200,000 words. For each subcorpus, I made an index and searched for 4-grams. The number of 4-gram types occurring in a set number of texts was then calculated, with the number of texts determined by the size of the subcorpus as in Biber and Barbieri's (2007) discussion (Table 4.9).

Subcorpus size (in words)	No. of texts set as minimum	No. of texts in subcorpus	No. of 4-gram types found
50k	3	23	24
50k retry	3	24	29
100k	4	40	48
100k retry	4	45	45
150k	5	74	40
150k retry	5	67	47
200k	6	105	63
200k retry	6	98	53

Table 4.9 Comparative normalization of types

The 50,000 word subcorpus (23 texts) produced a total of 24 four-gram types, each of which occurs in at least three texts. However, a second attempt (labelled 'retry') with a different 50,000 word corpus (24 texts) resulted in an increased number of types (29). Although at the 100k and 150k size, the number of types found is similar (ranging from 40 to 48), there is another change at 200k to 63 and 53 types. The findings overall show both different numbers of types for each corpus size, and a lack of consistency within searches, contrary to Biber and Barbieri's (2007) claim. It is more viable to count types in a corpus consisting of similar-length texts, such as ICLE, since the internal consistency of the corpus reduces the skewing effect experienced here. In a corpus of naturalistic texts, such as BAWE, statistical comparisons within the corpus are less reliable due to the internal variation of text lengths.

This section has examined the use of n-gram token and type counts as a means of comparing n-grams across datasets, concluding that while a case can be made for comparing n-gram token counts across corpora it is more difficult to make reasonable comparisons of n-gram types. Having established the parameters chosen for extracting and counting n-gram tokens, the next section discusses the structural and functional categorization of frequent 4-grams.

4.3.4 Analyzing 4-grams

In order to adequately categorize and describe findings from the potentially large dataset which can be produced from n-gram analysis it is necessary to set boundaries for the data, for example by considering just one size of n-gram. In their discussion of n-grams of different lengths, Stubbs and Barth (2007: 269) describe 2-grams (e.g. *at the*) as predominantly preposition-determiner combinations, 3-grams (e.g. *at the end*) as function and content word combinations (often containing delexicalized content words), and 4-grams as having 'clear differences in topic and function' between different registers (e.g. *at the end of*). Similarly, 4-grams are described by Cortes (2004: 401) as presenting 'a wider variety of structures and functions to analyze' when compared to 5-grams. Hence, in this study I focus my analysis on the most frequent 50 four-grams across the student corpora. The rationale for this is that in my corpora 4-grams are short enough to be sufficiently numerous yet long enough to be frequently contentful. This pragmatic decision of providing sufficient, yet not excessive, data for analysis is frequently adopted (e.g. Cortes, 2004; Hyland, 2008a,b; Scott and Tribble, 2006). Four-grams have been extensively employed in n-gram analysis and thus provide a basis for comparison with other studies (Cortes, 2004: 401).

Using the criteria established in 4.3.3 of 20 per million words (pmw) and across 5% of assignments, 4-grams from each of the four corpora were extracted; these n-grams were additionally checked to ensure they occur in the work of at least three students to reduce idiosyncrasies (a small number of n-grams from the Chinese corpora were omitted on this criterion). Examples which overlapped were discounted (e.g. *is due to the* occurs 43 times in Eng12 and *this is due to* occurs 44 times. Examination of the concordance lines revealed that 43 instances are part of the 5-gram *this is due to* the, so the smaller set of 4-grams was excluded from analysis. Partially subsumed 4-grams were added or reduced from the counts of longer types; for example, *the end of the* occurs 77 times in Eng12 and *at the end of* occurs 51 times. Thirty five instances of *the end of the* were found to occur within the 5-gram *at the end of the* and thus the 4-gram count for *at the end of* was reduced from 51 to 16 to avoid inflation of n-gram counts. Reduction of n-gram counts is rarely mentioned in the

literature, yet can have a distinct inflationary tendency on results (see Chen, 2009, for detailed description of her procedures in avoiding data inflation).

The 4-grams were then categorized structurally, using Biber et al.'s (1999) classification (see discussion in 3.5.1) and grouped functionally, using an adapted version of Hyland's (2008a,b) categorization system (discussed in 3.5.2). However, applying Hyland's functional categories did not prove straightforward as these assume that each type has a single primary metafunction and that all instances of the type can be classified in this way. To take one n-gram as an example, one of the most frequent 4-grams across both Engineering and Biology (and indeed, a frequent n-gram in English generally) is *at the end of*. This n-gram is particularly interesting since it has different uses in each discipline. In Biology, 11 out of 13 instances of *at the end of* are followed by a period of time, for example *at the end of + the trial/September/the survey period/1975* while the remaining two relate to a physical location. This contrasts with Engineering where just 15 out of 27 examples (from Chi-Engineering and Eng-Engineering) relate to a period of time and the remainder (12 out of 27) refer to a physical entity (*at the end of the pivoting rod, at the end of the beam*). In his work using a functional classification of 4-grams, Hyland does not specifically assign *at the end of* to any one functional category. However, the difference in usage outlined here suggests that it belongs in research-oriented 'location' (indicating time or place) in Biology (Hyland's examples include *at the beginning of, in the present study*), and in research-oriented 'description' for the Engineering cases where it describes an object (Hyland's examples include *the size of the, the top of the*). In this study I follow Hyland in keeping to a single functional categorization of each type, and classify all instances of *at the end of* as 'research-location' as this is the dominant category for disciplines overall. This decision is taken as a pragmatic solution to a complex problem since the alternative of classifying each token individually is not workable on a large scale. The most frequent fifty 4-grams are discussed in the consideration of writing across the year groups (Chapter 6).

4.4 Text level analysis

This final section focuses on procedures carried out on whole texts which are used to complement the corpus linguistic analyses. Analysis of pairs of assignments in three disciplines was carried out in order to compare students' use of visuals and lists in their responses to the same questions; this is reported on in 7.3. Also in Chapter 7, selected notes taken from the BAWE lecturer interviews are used for information pertaining to tutors' preferences for visuals and lists in assignments.

Multimodal comparison of assignments

Analysis of pairs of assignments in Biology, Economics and Engineering was carried out to examine whole assignments in detail and consider how visuals and lists are used by students. Since BAWE texts were collected at just four universities (Warwick, Reading, Oxford Brookes and also at Coventry towards the end of the project), there are a small number of texts by Chinese and by English students from the same university, same module and which answer the same assignment question. Pairs of texts in Biology, Economics and Engineering (each from one Chinese student and one English student) were found.

The analysis in this section draws on recent work in multimodality, in viewing assignments as comprising multiple modes and in analyzing the way visuals are used alongside the traditionally privileged mode of language (e.g. work by Archer, 2006; Jewitt, 2009; Kress, 2009; Kress and van Leeuwen, 1996, 2001; Van Leeuwen, 2005). A multimodal approach to the analysis of text considers all the 'semiotic resources' available, that is, the 'actions, materials and artefacts we use for communicative purposes' (Van Leeuwen, 2005: 285). These semiotic resources each possess a 'meaning potential, based on their past uses, and a set of affordances based on their possible uses' (p.285).Thus, instead of a concentration solely on language, a multimodal approach considers features such as the use of visuals in terms of their size, colour and integration with the text; the use of bulleted or numbered lists as a presentation device; and the layout of both visuals and text on the page. For the multimodal analysis in this section, pairs of texts from Biology, Economics and Engineering were found which in each case answer the same assignment question. The Biology texts

illustrate the different use of visuals in each L1 group, the Economics texts show how answers to a series of questions can be given in list format or as connected prose, and the Engineering texts indicate different approaches to layout.

For the two Biology assignments, counts were carried out of the number of tables, figures, and the size of these in proportion to the whole text (expressed as a percentage). Comments are also made on layout (continuous running text across the page or writing presented in two columns), the use of colour, text wrapping round visuals, changes in font, and the inclusion of lengthy captions for figures. In Economics, the use of bulleted lists and connected prose are compared, with discussion centring on the effect of the different assignment layouts. In Engineering, again the number and type of lists and the effect of the overall layout are compared.

The assignments, while few in number, appear to be typical of each student group and discipline within the dataset for this study and may provide insights into how Chinese and English students employ resources differently in answering the same questions.

Insights from BAWE interview data

The whole text discussion of a small number of assignments in 7.3 also draws on semi-structured interview data from the BAWE project. These interviews were carried out in a range of disciplines under the project ‘An Investigation of Genres of Assessed Writing in British Higher Education' during compilation of the BAWE corpus. A total of 58 lecturers from the three main participating universities (Oxford Brookes, Reading and Warwick) were interviewed. Topics covered include the types of assignment set, the role of written assignments in the department, and the aspects of student writing which are most valued by tutors (Leedham, 2009; Nesi and Gardner, 2006). As a researcher on BAWE I took part in almost one-third of the interviews and have access to data from all 58; section 7.3.4 draws on relevant comments by lecturers in Biology, Economics and Engineering to provide insights into lecturers’ perceptions of student writing. All lecturer quotations taken from this interview data are anonymized and are ascribed to the relevant discipline.

The non-corpus linguistic procedures described in this section are carried out in the whole text investigation in 7.3. While this analysis concerns only a small number of assignments, however, it is useful in providing a different perspective on the writing.

4.5 Chapter summary

Chapter 4 began by detailing the process involved in compiling and analyzing the datasets from BAWE and the additionally-collected assignments. From the 245 texts submitted to BAWE by Chinese students and the 99 additional assignments, the data was narrowed to the 146 texts (279,695 words) from undergraduate students whose secondary education was mainly in their home country. The resulting corpus (Chi123) contains writing from 12 disciplines, and the comparative corpus of Eng123 contains 175 texts (458,782 words) from the same 12 disciplines. The writing is investigated mainly through the corpus linguistic keyword procedure, using WordSmith Tools (v.5), and taking into account recommendations regarding reference corpora from Berber Sardinha (2004), Rayson (2008a), and (for Chapter 7) Baker (2004). The keywords are then followed up through exploration of concordance lines, collocate lists and dispersion plots to explore the context in which they appear. In addition to keyword and key n-gram analysis, overall text characteristics are given to provide an overall description of the corpora, the number of n-gram tokens in each student corpus is compared, and the most frequent 50 four-grams are analyzed both structurally and functionally. These procedures build a picture of the two student corpora (in the case of the text characteristics), and provide a means of extracting and analyzing n-grams (in the case of the keyword analysis and categorization of frequent 4-grams) in order to analyse aspects of student writing.

CHAPTER 5 FEATURES OF CHINESE STUDENTS' WRITING IN THE CORPUS

5.1 Introduction

The literature review in Chapter 2 surveyed the available empirical studies on Chinese students' undergraduate writing in the UK, and broadened the review to consider learner corpus studies of all NNS writers. Findings from these studies point to NNS' high use of particular lexical items and chunks, including informal or 'speech-like' items (e.g. Lee and Chen, 2009; Paquot, 2010) and particular connectors (e.g. Bolton et al., 2002; Milton, 1999), as well as high use of first and second person pronouns (e.g. Lu, 2002; Petch-Tyson, 1998). However, most of the reviewed studies (in 2.3) used data from learner corpora comprising collections of 500-word, argumentative essays, and it is unclear how far findings can be extended to the undergraduate assignments in BAWE (described in 4.2.1).

This chapter reports findings from my comparison of the two student corpora overall (Chi123 and Eng123) with the aim of answering research question 1:

RQ 1: What are the distinguishing characteristics of writing in English in a corpus of Chinese undergraduates' assignments in the UK?

In this chapter the Chinese student corpus (146 texts, 279,695 words) is interrogated through various searches to uncover features of the writing which distinguish it from the larger corpus of writing by English undergraduate students (611 texts, 1,335,676 words). First, overall text characteristics of mean assignment length, mean sentence length and mean word length are discussed as these provide a broad overview of the writing produced by the two student groups. The main means of uncovering variations in the data is through keyword searches which provide corpus-driven insights into potential differences. The keywords are categorized and explored through concordance lines, collocate lists and dispersion plots (as described in 4.3.2). In this chapter, the corpora are explored as a whole, with all three year groups and twelve disciplines remaining undifferentiated. However, each

search is checked to ensure it contains a spread across texts from different year groups and from a minimum of three disciplines, and thus is representative of the texts in each corpus.

5.2 Text characteristics

Measures of overall text characteristics are often used in studies of student writing (e.g. Engber, 1995; Gardner, 2008; Grant and Ginther, 2000; Jarvis et al., 2003) and provide an indication of broad points of comparison between two corpora. This section provides the average assignment lengths, sentence lengths, and word lengths of texts in the two datasets; these were derived through the WordSmith Tools 'statistics' feature on the Wordlist tool (as described in 4.3.1) and through Excel (for the mean assignment length and for calculations of significance). These statistics indicate that the mean assignment length (MAL) and mean sentence length (MSL) are significantly lower in Chi123 than in Eng123, but that the mean word length (MWL) is significantly greater in Chi123 (Table 5.1).

Statistic	**Chi123**	**Eng123**
Number of assignments	146	611
Number of tokens	279,695	1,335,676
Mean assignment length (MAL)	1916.42	2186.64*
[Standard Deviation]	[898.29]	[920.65]
Mean sentence length (MSL)	19.87	22.79***
[Standard Deviation]	[13.39]	[13.96]
Mean word length (MWL)	4.97*	4.91
[Standard Deviation]	[2.9]	[2.87]

Table 5.1 Descriptive statistics for Chi123 and Eng123 (significant difference using a z-test * p=.025, *** p=.001)

The standard deviation figures for each statistic are approximately the same for each student group, indicating similar distributions around the means (i.e. results are not skewed by a small number of students' texts). Each characteristic is discussed further below.

Mean assignment length

Examination of assignment lengths across different groupings indicates that the lower MAL in Chi123 is not confined to a particular year group, discipline, or genre, but holds true for all these comparable data groupings within the two corpora. One possible reason for Chinese students producing, on average, less writing than the English students for any given assignment is that writing in an L2 is more difficult and time-consuming than writing in L1. While this reason undoubtedly has validity, it is also the case that a lower wordcount is not necessarily a negative feature of texts. An assumption often made in second language writing research is that longer is better in terms of text length (e.g. Larsen-Freeman, 2006). While this belief may be warranted in the case of short test papers (e.g. Grant and Ginther's, 2000, study of timed essay tests), it is not always the case for high-scoring, untimed assignments. Concision is often a valued characteristic of academic writing, and verbose writing a deficiency (this point is made by several BAWE lecturer interviewees, and is discussed in 7.3.4). Clearly, in the case of these successful writers, the lower MAL is not associated with poorer quality work.

A further complicating factor affecting the MAL is that the assignment wordcounts are those provided by WordSmith Tools from the text files rather than being wordcounts from the original Word documents. Tables, figures, graphs and appendices are thus excluded from the wordcount as these are deleted (and the features tagged) before the text files are added to the corpus (cf. Gardner's, 2008, comparison of different ways of measuring assignment wordcounts in BAWE). Notably, all captions for visuals in BAWE are also omitted from the plain text files, thus words within captions are not included in the assignment wordcounts. Giving responses in the form of tables, figures, and so on can therefore result in a decreased wordcount in the treated files and this should be taken into account when considering the lower MAL in Chi123 (this point is discussed further in 5.6).

Mean sentence length and mean word length

The MSL is a measure of the average number of tokens per sentence; for Chi123 this is significantly lower than for Eng123 (p=.001), and holds true across sections of the corpus and across individuals' work. In Chi123 the MWL is significantly higher than for Eng123

(*p*=.025), and again is not skewed by a minority of texts. This difference between the MSL and MWL in the two corpora is counter-intuitive, as longer words and longer sentences might be expected to occur in the same texts through more sophisticated language use such as greater use of disciplinary terminology (longer words) and sentences consisting of multiple clauses (and thus longer). However measuring the correlation between the MSL and MWL across Chi123 and Eng123 revealed that this was not statistically significant (using a 2-tailed t-test, *p*=.01). Although the correlation is too slight to be significant, it is still of interest in that the low MSL and high MWL of Chi123 point to a difference between the corpora. I hypothesized that the Chinese students might be omitting short functional words: this would give both a shorter MSL (as there would be fewer words in the sentence) and a longer MWL (as there would be fewer short words). I compared the negative keywords for Chi123 and Eng123, calculated by WordSmith Tools, to determine whether any particular functional words were used more frequently in Eng123 than Chi123 (Table 5.2).

Key word	Freq.	%	RC Freq.	RC %	Keyness
that	2179	0.78	15527	1.16	-342
this	1611	0.58	12039	0.90	-322
would	407	0.15	3416	0.26	-134
be	2274	0.81	13940	1.04	-131
been	305	0.11	2706	0.20	-124
criminal	6		432	0.03	-122
women	37	0.01	768	0.06	-122
whilst	7		433	0.03	-117
these	357	0.13	2975	0.22	-114
have	745	0.27	5270	0.39	-111
within	138	0.05	1489	0.11	-105
being	159	0.06	1567	0.12	-92
land	5		325	0.02	-89
upon	37	0.01	612	0.05	-79
class	52	0.02	718	0.05	-75
must	138	0.05	1327	0.10	-74
organic	12		352	0.03	-70
a	5045	1.80	27300	2.04	-70
seen	90	0.03	972	0.07	-69

Table 5.2 Top 20 negative keywords in Chi123

From left to right, the columns in Table 5.2 give each keyword, its frequency in Chi123 and the percentage of the corpus this represents. This is followed by the frequency in the reference corpus (RC) (Eng123), the percentage in this corpus, and the keyness value (in

descending order of negative keyness). Several of the negative keywords in Table 5.2 are in fact five characters or longer (e.g. *criminal, women, within*); these satisfy the requirements of dispersion across year groups, disciplines and individual texts, but are more frequent in particular disciplines (e.g. *criminals* occurs mainly in Law, but also in Sociology and Computing). The negative keywords *be,* and *seen* occur in the negatively key 4-gram *can be seen in* (389 occurrences in Eng123, full details in Appendix D) and are almost all followed by a data items such as *figure, table, appendix.* For example (in Eng123)*:*

(1) …the loading scheme used in the FE Analysis *can be seen in* Figure 5.12. (3091i)[26] [27].

(2) … after a failed sweep. The whole program *can be seen in* Appendix B and… (0263g).

While these passivized n-grams help to explain the negative keywords *be* and *seen*, these items account for only a small portion of the lower MWL in Chi123. The most frequent negative keyword in Table 5.2 is the indefinite article, accounting for 2.04% of tokens in Eng123 and just 1.8% of tokens in Chi123. One conclusion which could be drawn from this difference is that the Chinese students use the definite or zero article more than the English students (cf. Chuang and Nesi, 2006). An alternative, or perhaps additional, explanation is that more of the Chinese writing occurs within lists, and that this writing is more likely to be abbreviated: one aspect of this abbreviation is article omission. Lists may also include higher use of numerals. The number of visuals and lists across the two corpora is explored in section 6.5.

Summary

This section has described the overall characteristics of the texts in each corpus. Chi123 texts have lower wordcounts, contain shorter sentences, yet have slightly longer mean wordlengths than Eng123; all of these differences are statistically significant. It was proposed that the lower MAL may be partially accounted for by the use of visuals as an alternative to prose in giving information within an assignment (in the case of shorter assignments). Since visuals (and their captions) are deleted, these are not included in the WordSmith wordcount.

26 Here and throughout the thesis, quotations from student writing are provided exactly as given by the student. Note that quotations are given either within inverted commas or as numbered extracts, while any linguistic item under discussion is given in italics. Lexical items or chunks which are discussed and are within quotations from student writing are thus italicized.

27 The alphanumeric sequence in brackets following each example is the identification (id) code for the text in BAWE or additionally-collected data (the latter id codes begin with a '7').

It was also proposed that the lower MSL yet higher MWL is in part due to the omission of short functional words such as the indefinite article (a negative keyword for Chi123), and that some of these omissions may be from abbreviated writing or the use of numerals within lists.

5.3 Keyword analysis

Keyness searches provide a corpus-driven way of comparing corpora and discovering which words and n-grams merit further investigation through more qualitative means (see 4.3.2 for discussion of the keywords procedure). In this study I searched for single keywords and for key n-grams of 2 to 5 words in Chi123 with Eng123 as an RC, using the search parameters of a minimum frequency threshold of 20 for keywords and 2-grams, and six for the (less frequent) 3 to 5-grams, using the log likelihood test with a *p* value of .000001. The resulting keywords were checked (through concordance lines) to eliminate keywords occurring in fewer than three disciplines, from only one year group (years 1/2 and year 3), or in writing from fewer than five students. From the list of extracted keywords (given in full in Appendix D), I placed items into groups which I devised through an iterative process of classifying and revising; three out of four of these groups or 'key categories' reflect findings identified in the research literature. This section examines the resulting four categories: informal items (e.g. *lots, a little bit*); connectors (e.g. *on the other hand, last but not least*); use of the first person plural (e.g. *we, we also need to*); and references to data or visuals within the text (e.g. *the figure, according to the*). On occasion, items in these categories overlap; for example, *besides* is discussed in the section on 'connectors' but is also placed in the category of 'informal items'. Where smaller keywords are subsumed or partially subsumed by longer ones, only the longer keyword is discussed (e.g. *the other hand* occurs 54 times as a key n-gram and *on the other hand* occurs 56 times). The keywords are then explored through concordance lines, collocate and dispersion lists. For each category, similar items are also explored; for example while only *lots* is a keyword, the similar item *a lot of* is also discussed as this is semantically similar though did not occur frequently enough to warrant inclusion as a keyword.

5.3.1 Informal items

In this section, I report how Biber et al.'s (1999) linguistic analysis in LGSWE was used to verify that my categorization of items as 'informal' can be supported through corpus data. Biber et al. based their descriptions on the 40 million word Longman Grammar of Spoken and Written English corpus; this comprises four main types of source text: conversation, fiction, news and academic prose. It might be expected that informal items in Chi123 would occur within less formal genres of assignments such as within reflective pieces of writing or within (often labelled) reflective sections of assignments. To check this, the genre classification for each text was checked (using Heuboeck et al., 2008); the co-text for each instance was also read to determine whether it occurs within a reflective section of an assignment (e.g. the final reflective paragraph of a case study in which the student comments on the experience of carrying out the study or of writing the assignment, these sections are often labelled as 'reflective').

The list of items categorized as informal provides (limited) evidence for the learner corpus literature claim that L2 writing is more informal than L1 writing (Table 5.5).

Keyword	Freq.	%	RC Freq.	R.C.%	Keyness
besides	49	0.02	9		125
lots	29	0.01	26		35
a little bit	8		0		28
lots of	24		24		27
last but not least	10		0		24

Table 5.5 Keywords containing informal items in Chi123
Key: RC Freq.= (raw) frequency of the item in the RC Eng123

It should be noted that the raw frequencies for these items are not large; excepting *besides* (49 occurrences), each keyword occurs under 30 times in Chi123 and thus does not indicate widespread usage across the majority of Chinese students' texts. Grouped together as a key category, however, the items point to a slight tendency towards informal language on the part of the Chinese students. *Besides,* and *last but not least* are categorized as both informal items and as connectors, and are explored in the next section; the other keywords (*lots [of], a little* bit) are explored below.

Lots (of)

Most occurrences of *lots* in the Chinese data are part of *lots of* (23 out of the 29 instances), a 2-gram described by Biber et al. (1999: 276) as 'characteristic of casual speech' rather than academic writing. The instances of *lots of* in Chi123 occur within otherwise formal co-text; for example:

(3) …BA has put *lots of* effort on advancing the airport and in-flight service and e-ticketing… (0123a).
(4) *Lots of* policies have been adopted to regulate the inappropriate behaviours of monopolies… (7008a).
(5) In order to design an appropriate brake speed, *lots of* things have to be considered… (0254i).

Notably, no instances of *lots of* in Chi123 occur within reflective writing, student letters or otherwise informal types of assignment. Similarly, in Eng123, only three instances of *lots of* occur within reflective sections of assignments. Example (6) below is from a reflective section, while (7) and (8) are from more formal pieces of writing (all taken from Eng123):

(6) During this meeting, *lots of* ideas can be floated, and only those… (6101d).
(7) …gravel base, covered with special silica sand, *lots of* post and rail fencing and… (6015f).
(8) …carnivorous fish means you need to feed them *lots of* wild fish, and in fish farms disease… (6035d).

The quantifiers *a lot of*, *plenty of* and *a couple of* are grouped together by Biber et al. (1999: 277) as items which 'most typically found in conversation or carry a strong overtone of casual speech when used'. However, none of these quantifiers occur significantly more frequently in Chi123 than Eng123, even when investigating further down in the keyword lists than the initial levels. Examples of *a lot of* from each student corpus are discussed below as an indication of similarities between the student corpora in the use of informal writing.

The 3-gram *a lot of* occurs 21 times in Chi123 and 75 times in Eng123 (not a significant difference). None of the collocates for *a lot of* are particularly informal; examples from Chi123 include 'a lot of health problems', 'a lot of trypsin activity', 'a lot of direct consumer interactions'; Eng123 examples include 'a lot of shareholders', 'a lot of potential business', 'a

lot of the fat'. A small number of instances in each corpus (approximately 5% in each) occur when the writer is giving their personal reflections and thus writing in a more conversational style:

(9) Secondly, as I did *a lot of* psychoanalysis in the past as part of my interest, I wanted to know myself better… (7014a, Chinese student).

(10) As I have never worked in a kitchen before, I was unsure of what to expect. I was expecting *a lot of* pressure particularly during service… (3103b, English student).

The majority of instances in both student corpora are used within otherwise formal pieces of writing, suggesting that this chunk is not viewed as particularly informal or inappropriate by students when writing assignments.

A little bit

Support for my categorization of *a little bit* as informal (or at least 'conversation-like') is given by Biber et al.'s (1999: 250) provision of examples for the 2-gram *bit of* from the conversation corpus (e.g. 'I watched *a bit of* television news'). Additionally, *a bit of* is discussed by Channell (1994: 104) within the category of 'vague quantifiers'. While a keyword in Chi123, however, *a little bit* occurs just eight times in this corpus (and zero times in Eng123) (Figure 5.1).

N	Concordance
1	the connection between GSM100T and PIC 18F452 is a little bit different. Because the serial port of modem is
2	those of the IBT and the conferences; however, there is a little bit different in the rate structure of the ILT. Since
3	and the probability of acceptance during sampling is a little bit higher than that of tightened inspection. By
4	was a great idea, but the title of our documentary will be a little bit long. "D'oeuvres" comes from France, it means
5	is slightly greater, so it seems that the process has a little bit more risk to produce products over the LSL than
6	denaturation of the serum proteins of the milk. It shows a little bit of browning because of Mailard reaction. There
7	City Centre Hotel. At that time, I found that this hotel is a little bit out of my expectation. There are three
8	were not match with them, and only the ductility was a little bit similar as the Appendix 1. So, the experiment

Figure 5.1 *a little bit* in Chi123

All eight occurrences in Figure 5.1 are from otherwise formal pieces of writing, with one exception (line 4); this line is from a piece of writing reflecting on the experience of working on a group documentary. To determine whether there are other, similar chunks, I searched for *bit* in both Chi123 and Eng123. Once lines with the sense of a computer *bit* had been removed, this yielded 20 occurrences from Chi123 and 25 from Eng123, rendering this a

keyword for Chi123 (p<.0001) (no instances of non-computer related *bits* were found in Chi123 and just 1 in Eng123 from a reflective section). A collocate search suggests that the most common chunk for both student groups is *a bit* followed by an adjective e.g. *a bit extreme/a bit high/a bit more difficult*. Two examples of this from Chi123 are:

(11) I found that the magnitude of filtered with ADC/DAC is *a bit lower* than the only… (6107a).

(12) … and only the ductility was a little bit similar as the Appendix 1. (0254c).

None of the 20 occurrences of *bit* in Chi123 are from reflective sections of writing, whereas one-third of the lines from Eng123 (8 of the 25) are from sections of the text which reflect on the experience of carrying out the assignment task, for example:

(13) The conclusion was also *a bit of a victim* in my editings, bringing it down to one small sentence for each of the areas of discussion (6101c).

(14) Trail pheromones pose *a bit of a* problem for ants though because they need to be long lasting but not to… (6035a).

A common pattern in Eng 123 (yet absent in Chi123) is *a bit of a + N* (n=6); for example '*a bit of an* issue', '*a bit of a* problem', '*a bit of a* shock', '*a bit of a* dog's breakfast' (though the intriguing use of the final idiomatic chunk in this list was a quotation from a Guardian article on the European Union, cited in a Law essay).

From the concordance lines it seems that both student groups use *a little bit,* and *a bit* in ways that have been considered informal within otherwise formal academic writing, though some of the instances in the English students' writing are within reflective sections.

So far, examination of this key category has provided slender evidence for Chinese students being more likely to employ informal items in otherwise formal writing. As there were very few keywords in this category meeting the minimum frequency threshold of 20 occurrences for a single word, I searched for keywords below this threshold. This search yielded the contraction *what's* (19 occurrences in Chi123, 4 in Eng123, keyness value of 47), and this led on to consideration of the broader category of contractions in student writing.

Contractions

The use of a contraction in academic writing is, according to Biber et al. (p.1129), highly unusual since these are 'strongly associated with the spoken language'. Most instances of the keyword *what's* occur within the connector *what's more* (n=13) and are discussed within the connectors category in the next section. Other instances of *what's* in Chi123 occur within a (repeated) citation of a report title (x 4) ('*What's* wrong with corporate social responsibility?'), a heading question which is presumably provided by the tutor ('Q1: What are free radicals and *what's* their relationship with antioxidants?'), with just one instance of *what's* occurs in the students' own language (it's likely that you will not know *what's* the meaning of LOLI, KUSO'). However, investigation of *what's* led me to explore whether there was a general tendency for either student group to use contracted verb forms.

Using the verb and negative contractions listed in Biber et al. (1999: 1128), I initially searched for all first, second and third person singular present and past tense contractions from the primary verbs *be* and *have* and from modal verbs. A search for the contracted person and auxiliaries (*I'm, you're, he's, she's, we're, they're*) yielded just two instances in Chi123 and seven in Eng123 (after removal of quotations) and these are not discussed further. Rarer modal contractions (*could've, may've, mayn't, might've, mightn't, should've, would've, wouldn't, shan't*) were included in the initial search but are omitted from the list as no occurrences were found in either corpus. The search list discussed in this section thus comprised the following:

> *aren't, can't, couldn't, doesn't, hadn't, haven't, hasn't, isn't, shouldn't, wasn't, weren't, won't*

The search revealed that as a group, these contracted verb forms are significantly more prevalent within Eng123 than in Chi123 (285 in Eng123, 32 instances in Chi123, p=.01), although none of the verb forms in isolation is a keyword. Very few of the instances of contracted verb forms in Eng123 are from reflective pieces of writing, but are used within otherwise formal, academic assignments, by a range of individuals and occur in a variety of disciplines, as illustrated in the examples below (all taken from Eng123):

(15) Politics is crucial for Lukes' argument as he believes A *doesn't* exercise power over B simply through having the greater ability to effect their own…(0004d).

(16) A point to note is that if the public do use the restaurant as a venue, then the members *can't* be excluded from the club house, … (3101g).

(17) When S-N curves *aren't* available for fatigue life predictions, the following assumptions are made in order to… (3091i).

(18) This stakeholder group *isn't* part of the firm itself and as such has little formal influence over the changes made. (0193d).

(19) As it happened, the system *wasn't* fully understood by the staff because Tiptree struggled to find the time to train them properly due to rigorous testing… (6106b).

It is unclear why the English students make greater use of this informal feature in their assignments compared to the Chinese students. Biber et al.'s (1999: 1128-1132) discussion of both verb and *not* contractions repeatedly points out the rarity of these in academic writing; however, the academic prose in the LGSWE comprises book extracts and research articles, rather than the more varied macro genre of student academic writing. The formality ascribed to professional academic writing by Biber et al. may simply be less applicable to successful student writing. It may also be the case that academic writing, or at least *student* academic writing, is becoming less formal than at the time the texts within Biber et al.'s study were written (the LGSWE was first conceived as a project in 1992 and completed in 1999). Perhaps a prohibition on contractions in academic writing is stated (e.g. in EAP texts and writing guides) more than it is actually observed in practice, hence the Chinese students avoid these forms (as they have spent longer studying EAP) while the English students embrace them. An additional (tentative) hypothesis is that the requirement for informal reflective writing at undergraduate level may have started to influence other genres of student writing, rendering contractions more acceptable, at least to the English students.

Section summary

The literature reviewed in 2.3 emphasized NNS students' tendency to use informal or 'speech-like' chunks. However, the data reviewed in this section is sparse and contains only a few key items in this category (*besides lots (of), a little bit, last but not least, last but not least*). Considering items below the minimum frequency threshold set for keywords (of 20

items) provided *what's* (found mainly within *what's more*) and gave rise to a search for contracted verb forms and contracted negatives. This search revealed that, although no individual items were key in the English corpus, the set of contractions as a whole were used significantly more frequently in Eng123. While some of these instances are from informal sections of writing (e.g. within the genres of empathy writing or reflective recounts in the narrative recount family, as classified in Heuboeck et al., 2008), most occurrences are within otherwise formal writing by English students. Taking the findings for informal writing as a whole, this study has provided only slim evidence for the literature finding that NNSs employ more informal lexical items and chunks. Moreover, the finding that NS students employ more contracted verb forms than NS students indicates that NS student writing also has informal aspects.

5.3.2 Connectors

The literature on NNS writing suggests that NNSs generally, and Chinese students in particular, favour particular connectors and that they use these repeatedly (e.g. Bolton et al., 2002; Hyland, 2008a; Lee and Chen, 2009; Milton, 1999). The term 'connectors' is used here to refer to lexical items which have a broadly textual function in connecting parts of the writing (these items are termed 'linking adverbials' in Biber et al., 1999: 875). This key category includes 12 n-grams (excluding shorter n-grams subsumed within these such as *on the other* from *on the other hand*) and two negative keywords key (i.e. items occurring significantly more frequently in Eng123 than Chi123). The keyword connectors given in Table 5.4 are either single words or relatively fixed, multi-word chunks, that is, there is little space for substituting any intermediary words (with the exception of *in the long run* where *long* can be replaced by *short/medium/long*, and *on the one hand/on the other hand*).

Keyword	Freq.	%	RCFreq.	R.C.%	Keyness
besides	49	0.02	9		125
nowadays	33	0.01	9		75
in other words	29	0.01	13		55
meanwhile	22		6		50
and so on	23		9		46
what's more	13		0		46
on the other hand	54	0.02	81		38
nevertheless	47	0.02	73		32
last but not least	10		0		24
at that time	14	9	22		14
in the long run	19		19		21
at the same time	35	0.01	65		18
Negative keyword					
however	395	0.14	2605	0.19	-39
therefore	281	0.10	2027	0.15	-47

Table 5.4 Key connectors in Chi123

This section focuses firstly on connectors overlapping with the key category of 'informal items' (*besides, what's more, last but not least*), then considers *on the other hand*, and finally examines the negative keywords.

Informal connectors

Three of the key items (*besides, what's more, last but not least*) were mentioned in the previous section as appearing informal for academic writing; these items occur within otherwise formal co-text. For example:

(20) *Besides*, IHG has a pension deficit of £172m in… (3018e).

(21) *What's more*, the location of double bonds in the chain can also change the nature of the fatty acid… (6081l).

(22) *Last but not least,* the appended PCA plot represents us… (6150c).

The connectors *besides* and *what's more* are also keywords in Lee and Chen's (2009) corpus of Linguistics writing by Chinese undergraduate students (with BAWE Linguistics writing as a RC) (2.3). As in Lee and Chen's study, both connectors are used in Chi123 with an additive function in the sense of *additionally* or *moreover*, for example:

(23) *What's more,* Butterworth filter will have a more linear phase response in … (6107a).

(24) *What's more,* the eating habit of adolescents changed a lot while the high… (6081k).

(25) … it is more likely that they will run into diminishing returns. *Besides,* human capital plays a vital role in production. (0080b).

(26) … (ROS) and free radicals in the body is an important factor leading to cancer. *Besides,* ROS appear to be involved at all stages of cancer development. (6150b).

(27) …they favour illegal immigration which makes things worse. *Besides that,* the neo-liberal economists have pointed out that there are potent… (7001b).

While the additive function is the standard one ascribed to *what's more*, Lee and Chen comment that it is more usual for *besides* to indicate that the ensuing point is one of subsidiary detail rather than a major addition to an argument. None of the Chi123 instances of *besides* or *what's more* are used within reflective sections of assignments, indicating that they are used as alternatives to more formal connectors. Similarly, the connector *last but not least* is used in formal writing, and indicates a concluding point in an argument in the same way that *in conclusion* or *finally* might be used:

(28) *Last but not least*, free radicals generated at sites of inflammation during infection can attack the host cell, leading to apoptosis and necrosis. (6150c).

(29) *Last but not least*, the market demand elasticity is also essential in incurring different levels of monopoly's economic inefficiency and welfare loss. (7008a).

This connector occurs only ten times in Chi123, but is salient on reading these assignments as there are zero instances of the 4-gram in the larger RC (Eng123).

On the other hand

On the other hand has been discussed in studies of NNS writing as a particularly highly-used sequence (e.g. Milton, 1999). This chunk is the most frequent connector used by the Chinese students in the study (54 occurrences), though not the most key, and is widely dispersed across texts, individuals and disciplines in Chi123. Examples include:

(30) *On the other hand,* Sir Peter Millett has argued extra-judicially that where the breach of trust involves an unauthorized act … (0410a).

(31) *On the other hand*, individuals with a strong achievement motive perceive accomplishment as an ends… (0271c).

The accompanying connector *on the one hand* occurs far less frequently (just once in Chi123, ten occurrences in Eng123). For Chinese students, the 4-gram *on the other hand*

may be frequently used as it is regarded as equivalent to a Mandarin expression meaning 'two sides of a coin', and is seen as having a more strongly contrastive meaning than the popular Eng123 connector *however*[28]. An additional possibility for the high use of *on the other hand* is North's (2003: 336) suggestion that some students may choose a longer chunk to increase the word count of their assignment. This particular 4-gram is frequently-used in all academic writing (Hyland, 2008b) and the higher use in Chi123 indicates a slight preference for this over alternatives such as *however*.

Negative key connectors

The two connectors which occur significantly more frequently in Eng123 (*however* and *therefore*) are two of the most common single 'linking adverbials' in academic prose (Biber et al., 1999: 887). These two items are particularly well-dispersed across the English texts with *however* occurring in 516 assignments and *therefore* in 460 (of the 611 in Eng123). Searches for other single-word connectors commonly used in academic writing such as *thus, then, furthermore, hence, nevertheless* (from Biber et al., 1999: 887) did not result in any significant difference between the student groups, that is, only the two single-word connectors *however* and *therefore* appear to be key in Eng123.

As well as being used significantly more frequently in Eng123 than Chi123, *however* and *therefore* are also more likely to be used in varied positions by the English students (35% of occurrences of *however* and 69% for *therefore* are *not* sentence-initial in Eng123; the figures are 12% and 40% respectively in Chi123). Examples of these connectors used in medial position in sentences in Eng123 are given below:

(32) Liverpool etc following the Brixton riot; *however* as Waddington (1992) argues… (0408d).
(33) Calves are *therefore* born effectively with no immunity… (6015j).

In Chi123, they are more typically used sentence-initially:

(34) *However,* it is doubtful whether other non-state norms apart from codifications… (0410e).
(35) *Therefore,* as the Bretton Woods system evolved, the… (0197a).

28 These points were suggested to me by three Chinese informants (Guozhi Cai, Liang Wang and Yu [Torri] Wang).

High use of connectors in sentence-initial position in NNSs' writing has been noted in previous studies (e.g. Milton's, 1999, study of Hong Kong Chinese students' writing), and has the effect of foregrounding the connector as a marked textual theme (Halliday, 1994). Textbooks in the PRC, however, commonly present connectors in sentence-initial position, whether through sentence-level exercises or within short pieces of writing, and this is likely to influence the way in which students use connectors; this point is taken further in the textbook analysis in 6.5.

Summary

The evidence reported in this section supports the suggestion in the literature that Chinese students as a group favour particular connectors (the 12 appearing as keywords), and that some of these connectors appear informal for academic writing (though only three of the 12 are informal). Moreover, the Chinese students make significantly lower usage of the single-word connectors *however* and *therefore* than the English students. Variation in the use of connectors across year groups is investigated further in 6.4; connectors are also the focus of a small study of textbooks and model examination answers examining the possible influence of textbooks in priming Chinese students to use particular informal chunks (6.5).

5.3.3 First person pronouns

First person pronouns have recently been the focus of research into both student and professional academic writing with several studies examining the role they play (e.g. Harwood, 2005; Hyland, 2002; Luzón, 2009; Martinez, 2005; Tang and John, 1999). The presence of first person pronouns in academic writing in general is described by Luzón (2009: 193) as 'refuting the traditional view that this type of discourse is impersonal and objective' since these pronouns are often viewed as indicating a high degree of authorial involvement in writing. In the literature on NNS student writing (reviewed in 2.3), the use of first and second person pronouns are often described as 'overused' in comparison with 'expert' writing (e.g. Cobb, 2003; Lee and Chen, 2009; Lu, 2002; McCrostie, 2008; Petch-Tyson, 1998). The literature is not in complete agreement, however, as Hyland's (2002) study found that his Hong Kong Chinese students employed first and second person pronouns less frequently than professional academic writers. Of particular relevance to this

study is Lee and Chen's (2009) work on Chinese undergraduate writers (using a corpus of Linguistics dissertations written in PRC universities). Lee and Chen found that the first person plural was employed significantly more frequently by the Chinese students than in their RC of English undergraduates in Linguistics (extracted from BAWE), and view this as 'overuse'. This section reports on the high use of *we* in Chi123, in comparison with Eng123.

Preference for *we*

In this study, the first person pronoun *we* was retrieved as both a single keyword and as part of 2-grams and a 4-gram in Chi123 (Table 5.6).

Keyword	Freq.	%	RC Freq.	R.C.%	Keyness
we	591	0.21	1411	0.11	180
we will	40	0.01	42		43
we can	86	0.03	212	0.02	24
we could	25		30		23
we need	22		26		21
we also need to	6		0		21
Negative keywords					
they	612	0.05	3694	0.28	-30
their	536	0.04	3286	0.25	-31
his	111		993	0.0	-46
that they	46	0.02	467	0.03	-29
as they	29	0.01	311	0.02	-22

Table 5.6 Keywords containing first person plural in Chi123

Table 5.6 indicates that while *we* is a positive keyword in Chi123, some third person pronouns (including possessive pronouns) are negative keywords (i.e. *they, their, his, that they, as they*). To examine pronoun use more broadly I compared the use of first, second, and third person pronouns across the two corpora, conducting searches for person, possessive and reflexive pronouns under groupings according to Biber et al.'s (1999: 328) classification (Table 5.7).

pmw	Chi123	Eng123
***we**, us, our, ours, ourselves*	2921****	1658
***I,** me, my, mine, myself*	1527	1932****
***you,** your, yours, yourself, yourselves*	372	436
***she**, her, hers, herself*	272	514****
***he,** his, himself*	887	1700****
***it,** its, itself*	9042	9775***
***they,** them, their, theirs, themselves*	4977	6390****

Table 5.7 Pronoun use in the two corpora (***p<.001; ****p<.0001)[29]

Table 5.7 reveals that there is a disparity between Chi123 and Eng123 in several pronoun groupings (values normalized to occurrences per million). First person plural pronouns are more common in Chi123, while first person singular pronouns, third person singular and third person plural pronouns are significantly more common in Eng123. No significant difference was found in the use of second person pronouns in the two corpora. These findings contrast with the majority of studies in the learner corpus literature reported in 2.3 which found that L2 writers make greater use of all first and second person pronouns.

The remainder of this section focuses on the discrepancy in the use of first person singular and plural pronouns, since these are most closely linked with authorial presence in the writing and have provoked some discord in the literature (e.g. between Petch-Tyson's, 1998, claim of high use in student writing and Hyland's, 2002, claim of low use). Analysis is confined to the personal pronouns *we* and *I*; the next section categorizes instances of these to investigate differences in usage.

29 Unless otherwise indicated, all significance figures are calculated using Rayson's Log Likelihood calculator: http://ucrel.lancs.ac.uk/llwizard.html.

* 95th percentile; 5% level; p<.05; critical value = 3.84
** 99th percentile; 1% level; p<.01; critical value = 6.63
*** 99.9th percentile; 0.1% level; p<.001; critical value = 10.83
**** 99.99th percentile; 0.01% level; p<.0001; critical value = 15.13

Classifying first person pronouns

Studies investigating first person pronoun use frequently devise functional taxonomies in order to group pronoun use (e.g. Harwood, 2005, 2009; Hewings and Coffin, 2007; Mayor, 2006; Rai, 2008; Tang and John, 1999). To investigate the different uses of *we* and *I* in this study, I first retrieved a random 100 lines of each pronoun from each student corpus, then classified the instances according to primary function. Five categories were iteratively developed, based closely on those employed by Tang and John (1999) as these were devised for student writing and, for the most part, fulfilled the identified purposes of the pronouns in the texts. However, Tang and John's study was limited to 27 texts from a single genre, (and just 92 instances of *we/I*), thus some deviations from their classification were needed. One important difference is that I distinguish between *we* and *I* in the classification; this also allowed me to merge some of Tang and John's categories as pairs of categories employed *we* in one grouping and *I* in the other as, for example, the 'representative' category excludes use of *I* since inclusive *we* is required in order to include writer and reader. Additionally, I have added a 'reflective' use, as this purpose was salient in some of the texts, but was not a discrete category in Tang and John's study (cf. Braine and McNaught, 2007). The resulting five categories are *we/I* as representative, as guide, as recounter, as opinion-holder, and as reflecter. The categories are discussed and exemplified below:

1. **Representative.**

The first category identified was the use of a first person pronoun to refer to a group; this could be humans or society in general (e.g. 'Despite this, it can be argued that *we* are defying natural selection') or a smaller group of members of a discipline community (e.g. 'There are still large areas of the bacterium's biology *we* don't understand'). As this use refers to multiple persons either inclusively (i.e. including reader and writer) or exclusively (i.e. the writer and other people but not the reader), this function excludes the first person singular.

2. Guide

In this category the writer is guiding the reader through the assignment (e.g. 'In this section, *we* look at common law damages and equitable compensation'; *'I* will discuss the policies in 4 areas'). Tang and John distinguish between 'guide' and 'architect', describing the former role as the writer appearing to stand apart from the writing, whereas the architect signposts the writing for the reader. In practice, these roles are difficult to distinguish and I have chosen to amalgamate them in this taxonomy.

3. Recounter

We or *I* can be used to recount (i.e. describe the procedures carried out within the task of an assignment) and is often followed by past tense verbs (e.g. 'firstly *we* measured out an area', '*I* have also calculated these factors'). Use of pronouns to recount tends to occur in sequence as a student reports one procedure after another.

4. Opinion-holder

The fourth category is opinion-giving (e.g., '*I* personally find the use of motivational skills to be…', 'Based on our sensitivity report we believe that the suggested mix'). This merges Tang and Johns' categories of 'opinion-holder' and 'originator' as again, the two are difficult to distinguish. In their view, the former entails the expression of an opinion, usually occurring with a mental process (cf. Halliday, 1994), whereas the latter involves the writer presenting or signalling new 'ideas or knowledge claims' (Tang and John, 1999: 29). Originating knowledge is thus 'the most powerful role' (p.29) that a writer can portray, and is comparatively infrequent in student writing when compared to professional academic writing.

5. Reflecter

Finally, the new category of *we/I* as reflecter was added in order to distinguish pronouns used when students think about the process of carrying out the work for the assignment or reflect on their future careers (e.g. 'had the team taken advantage of this earlier, *we* may have performed even better on the Upper Level pitch', 'also, *I* remember *I* could not wait to

finish off everyday working…'). Similar 'reflecter' categories are also used by Thompson (2009) and Rai (2008).

From each student corpus, 100 concordance lines containing each first person pronoun were categorized. Some concordance lines appeared to satisfy multiple functions, and additional context was needed to determine a primary category (Table 5.8).

		Representative	Guide	Recounter	Opinion-holder	Reflecter
we	Chi123	27	26	41	3	3
	Eng123	45	20	28	2	5
I	Chi123	0	19	24	18	39
	Eng123	0	18	31	12	39

Table 5.8 Classification of functions of *we* and *I* in 100 random lines

The proportions of each pronoun in each category are roughly comparable for the two student groups. For Chi123 and Eng123, the dominant functions of *we* are as a representative of a group of people (particularly in Eng123), as guide through the writing, and as recounter of the procedures followed (particularly in Chi123). For *I,* the most highly used function for both student groups is to reflect on the writing and/or the experience of carrying out the work for the assignment.

The next two sections discuss the greater use of *we* to recount in Chi123 (both as a proportion of the categorized pronouns in Table 5.8 and in overall counts), and the high use of *I* to reflect in Eng123 (*I* as reflecter is similar across the student groups in terms of proportion of the classification in Table 5.8, but greater in overall counts in Eng123).

Greater use of *we* to recount in Chi123

The dominant function for *we* in Chi123 is the recounting of methodological procedures. For example:

(36) For the following statistical analyses, *we* used values of sample pairwise differences (6215f).

(37) In those experiments, *we* determined the density and specific gravity of a range of food by measuring the bulk density of solid and using specific gravity bottle to determine the particle density (6081c).

(38) However, before *we* start the distillation process, *we* need to add anti-bumping granules to ensure even heating inside the distillation flask (6008g).

Even though all the assignments selected for this study are written by individual students some of the experiments or studies were carried out as groupwork, rendering some uses of *we* to recount procedures congruent with the multiple actors involved. It is often difficult to know whether *we* is exclusive and refers to the multiple participants in the study, or whether it is inclusive and refers to the reader and writer as members of a disciplinary community. From reading assignments extracts from both student groups in which methodological procedures are recounted, it seems that while both groups use *I* and *we* to recount, the Chinese group are more likely to use *we.* It seems unlikely that the Chinese students in the study have carried out more groupwork than the English students, and more plausible that the former simply favour the use of *we* over *I.*

Greater use of *I* to reflect in Eng123

For both student groups, the dominant function of *I* is to reflect (39% of each set of 100 categorized lines). Common n-grams for *I* in Chi123 include *I feel that* (40 occurrences), *I believe that* (21), *I have learnt* (18), *I was able to* (15), all of which are indicative of reflective writing. To explore the dispersion of *I* within assignments I generated plot dispersion charts for the ten assignments with the highest number of 'hits' (or instances of use) for *I* (Figure 5.2).

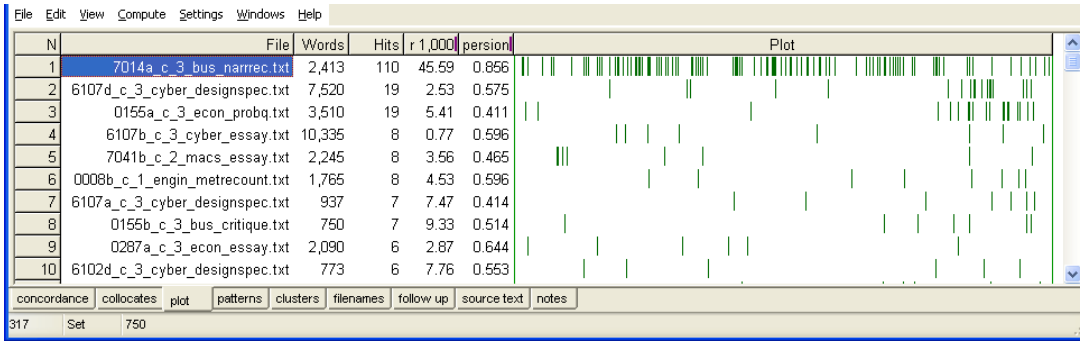


Figure 5.2 Plot dispersion for *I* in Chi123

For each of the ten assignments with highest use of *I,* Figure 5.2 shows the file name, number of words, number of hits, hits per 1,000 words, and dispersion across the assignment. The left hand side of each row under 'plot' represents the beginning of the assignment and the right hand side the end; each vertical line in the plot area represents an instance of the feature. The ordering of assignments is in terms of the number of hits, with the highest number at the top (110) going down to six at the bottom of the figure. The plot dispersion reveals that a single text accounts for 110 of the 317 instances of *I* in Chi123. Text 7014a is a year 3 Business assignment entitled 'Self Awareness', in which the student describes how they feel about a possible future career in Business:

(39) When *I* began this 'Self awareness' module*, I* learnt *I* am psychologically and physically boundary less, but prefer to stay in the same organization; also, *I* am self directed and actively manage my career in line with personal values which plays an important role of choosing employment (7014a).

Excepting this outlying assignment, the use of *I* is restrained in Chi123 with no other text reaching more than 19 hits.

The plot dispersion indicates that use of the first person singular is spread throughout these assignments, with occasional 'burstiness' or repeated occurrences close together (Katz, 1996: 15). Most of these bursts occur at the end of the text and result from reflective sections in which the student comments on the process of carrying out work for their assignment:

(40) In the process of trying different approaches, *I* gained a better understanding both on industry and academy (6107d).

(41) During the process of filming, *I* did not only review the academic knowledge from the lessons and books, but also… (7040c).

The prolific use throughout text 7014a and the small number of end bursts shown in Figure 5.2 account for most instances of the reflective use of *I* in Chi123. In contrast, 'end bursts' of *I* denoting reflective sections of the assignment are more widespread throughout the Eng123 corpus, as indicated in the plot dispersion for this corpus (Figure 5.3).

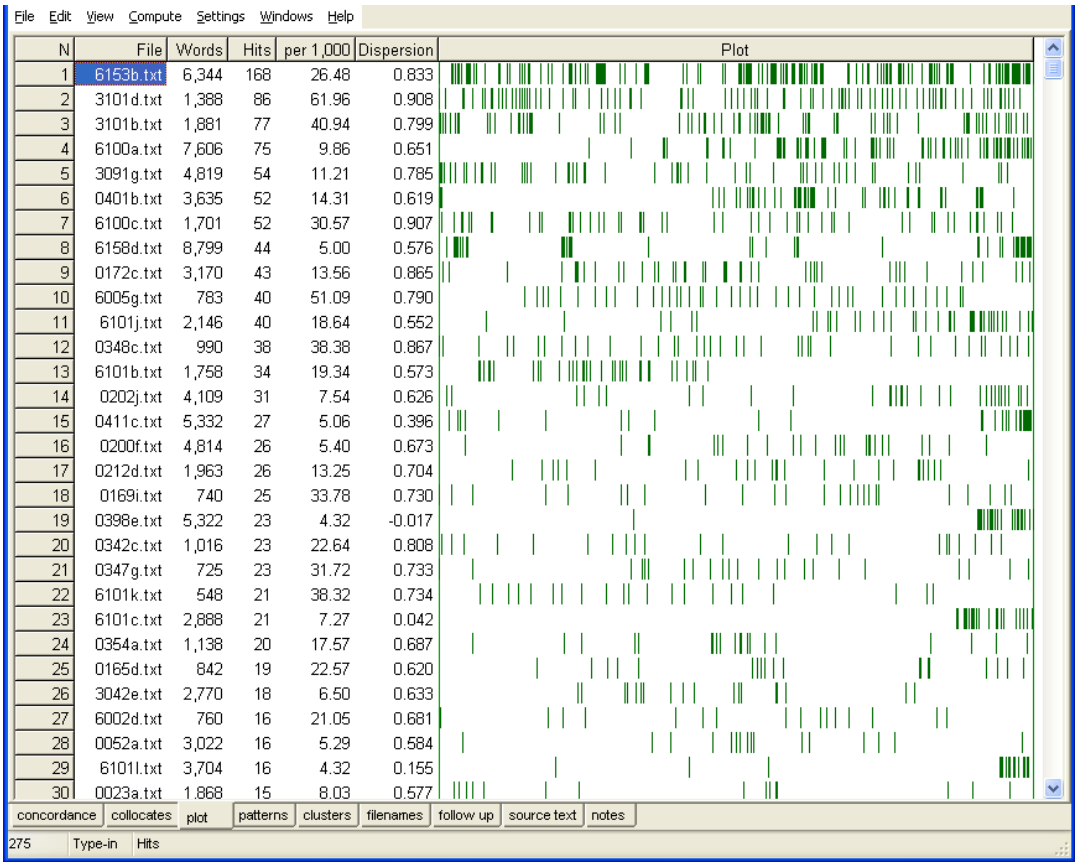


Figure 5.3 Plot dispersion for *I* in Eng123

The assignments in Figure 5.3 again fall into two groups. In some, the 'hits' for *I* are reasonably constant throughout the writing. Those with high numbers of hits throughout the assignment tend to be from the genre group classified in Heuboeck et al. (2008) as 'reflective recounts' from the genre family 'narrative recounts' in which the writer is discussing what they thought and how they felt. The examples below are from reflective

writing within an Engineering module called 'Starting and Running a Business' and from a Business module entitled 'Project Management' respectively:

(42) *I* decided to stand down as group leader at this point because Member A had emerged as a natural group leader and *I* felt it would make more sense to allow him to excel in this role...(0348c).

(43) Whilst the network planning analysis proved useful for overall project timing, *I* felt a need for a more visual, short-term means of monitoring task completion…(0172c).

In each of these assignments the student writer is foregrounded throughout the text as they describe their feelings about carrying out the work and writing the assignment.

The second grouping in the plot dispersion in Figure 5.3 is texts with 'end bursts' of *I*, indicated by few vertical lines in the text and a cluster of lines on the right hand side (e.g. rows 8, 15, 19). Reading the PDF files[30] shows that these 'end bursts' are due to a more personal, reflective stretch of writing occurring after the conclusion to the assignment. End bursts of *I* are not confined to any one discipline or genre family, occurring in, for example, Law essays, Business case studies, Cybernetics design specifications. The sample text below is headed 'self-reflection task' and is taken from the end of a Computing assignment in which the student reflects on the experience of writing software coding:

(44) As *I* got into this piece of work, *I* enjoyed doing it more than *I* expected. There's a great moment on each question where you press enter after putting in a lot of code or chasing a bug and it all just works. (6101l).

Similarly, the example below is from a student's end reflection in an essay entitled 'Social, Legal and Ethical Aspects of Science and Engineering':

(45) *I* always find word limits difficult. *I* either run out of steam before the end and am faced, as *I* am now, with another 400 words and no ideas, or *I* have so much to cram in that *I* have to leave swathes out. (6101c).

Both these samples seem to be required sections for the students' assessed work. It is difficult to know how far such reflective sections echo writers' feelings, since students know this is going to be assessed by a tutor and that the remainder of the writing may be read in

30 The original MicroSoft Word documents submitted by students are converted to PDF (portable document format) and provided to BAWE researchers on request.

the light of their reflections (see Nesi and Gardner, forthcoming, 2011, for fuller discussion on the use of reflective writing in undergraduate assignments in the BAWE corpus).

Summary

The analysis of first person pronouns in this section has focused on the use of *I* in reflective recounts and in reflective 'end bursts' of texts since this helps to account for the higher use of this pronoun in Eng123. It seems that Chinese students are less likely to write reflective pieces (or at least that they are less likely to gain good marks and to submit this writing to the BAWE corpus), and that they less commonly write end reflective pieces (or again receive lower marks for writing containing these). The corpora can only reveal findings from the assignments it contains, and I can only speculate as to the reasons for the absence of particular writing styles.

5.3.4 References to data and visuals

The final key category to be considered contains the largest group of keywords from Appendix D: references to data and visuals (Table 5.9). This category is also the only one which has not been predicted by the literature.

The keywords include numbers (denoted collectively in WordSmith by the hash symbol [#]), formulae (all mathematical, chemical or other formulae are replaced by the word FORMULA in BAWE tagging), and references or directives to data items (e.g. *according to the* + *figure/appendix/equation, refer to (the)* + *figure/table* + [number]).

Numbers (#)

The greater use of numbers in Chi123 was also followed up through concordance line searches and seems to be in part due to the use of numbered lists in the Chinese corpus (e.g. 'There are 3 generic ways of changing the structure of a market: 1. building a new or modified set of players in a market, 2. eliminating players in a market, 3…'). Numbers are additionally used to label tables and figures (*Table 4, Figure 3*); within equations; in references (footnotes are not deleted in BAWE and many contain dates and page numbers); and within percentages and other data.

Keyword	Freq.	%	RC Freq.	R.C.%	Keyness
#	10,541	3.77	44704	3.35	121
FORMULA	832	0.30	2,710	0.20	87
according to	141	0.05	242	0.02	82
as below	23		0		81
according to the	73	0.03	77		77
figure	287	0.10	790	0.06	58
the appendix	40	0.01	43		42
refer to the	29	0.01	22		40
the figure	22		12		37
based on the	71	0.03	134	0.01	35
eq	22		14		32
refer to	39	0.01	55		30
standard deviation	25		23		30
refer	42	0.02	64		29
referring to	24		23		28
the equation	43	0.02	74		25
according to the equation	7		0		25
illustrated in	21		20		24
FORMULA and	43	0.02	82		21
FORMULA is	40	0.01	74		21
in the appendix	22		26		21

Table 5.9 Keywords containing references to data in Chi123

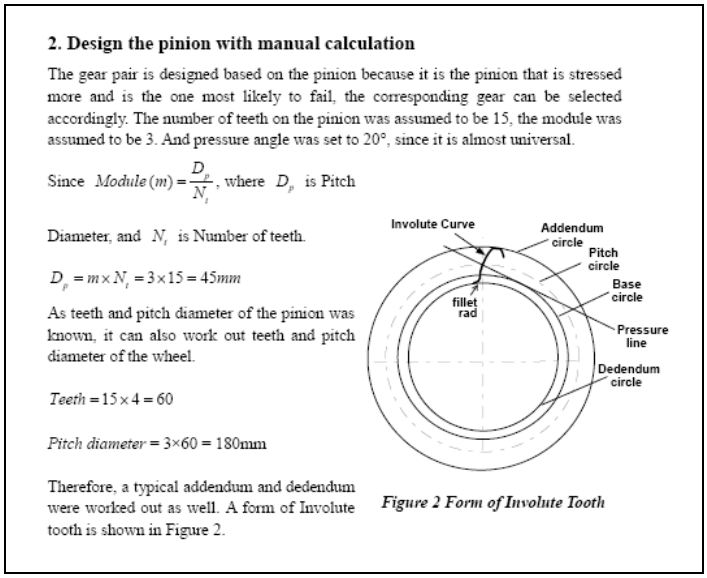
Formulae

Formulae appear more frequently in Chi123, and are often linked together as an argument.

For example (shown as it appears in the tagged text version):

(46) Since FORMULA , where FORMULA is Pitch Diameter, and FORMULA is Number of teeth 0254j).

The same sentence is given in Figure 5.4 as a screenshot from the PDF file with the formulae in full and the surrounding co-text and figure:



2. Design the pinion with manual calculation

The gear pair is designed based on the pinion because it is the pinion that is stressed more and is the one most likely to fail, the corresponding gear can be selected accordingly. The number of teeth on the pinion was assumed to be 15, the module was assumed to be 3. And pressure angle was set to 20°, since it is almost universal.

Since $Module\,(m) = \frac{D_p}{N_t}$, where D_p is Pitch Diameter, and N_t is Number of teeth.

$D_p = m \times N_t = 3 \times 15 = 45mm$

As teeth and pitch diameter of the pinion was known, it can also work out teeth and pitch diameter of the wheel.

$Teeth = 15 \times 4 = 60$

$Pitch\ diameter = 3 \times 60 = 180mm$

Therefore, a typical addendum and dedendum were worked out as well. A form of Involute tooth is shown in Figure 2.

Figure 2 Form of Involute Tooth

Figure 5.4 Extract from 0254j showing integration of formulae

The effects of the text wrapping round the figure and the length of the (italicized) formulae linked by connectors are not apparent in the tagged version of the text. For a more complete picture of how the student uses words, formulae and figures, it is necessary to read the PDF document. The interconnection of prose and formulae, and prose and figures is further explored in the whole text analysis of pairs of assignments in 7.3.

Tables and figures

A greater inclusion of tables and figures by the Chinese students was hypothesized from the presence of keywords such as *refer, figure* and *according to,* many of which are contained within (sometimes implicit) directives to the reader. For example:

(47) According to the program and *refer* to the *figure* 4.1.1, it is easy to find… (6107d).

(48) As shown in *Figure* 3, IHG even shows a better performance than… (3018e).

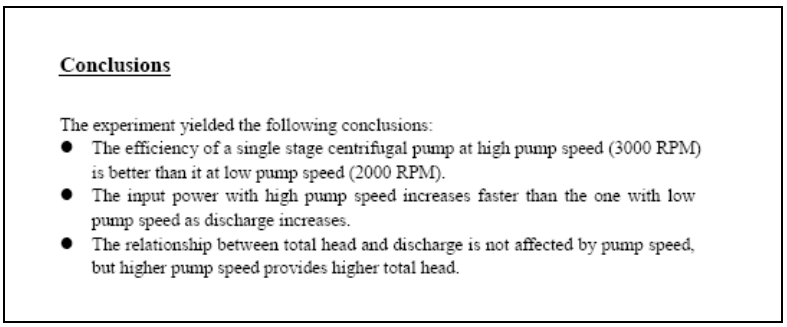
(49) According to the 3 sets of data calculated above… (6150d).

Counts of visuals and lists

The existence of frequent references to visuals does not in itself mean the Chinese students use more of these features in their assignments than the English students as it could be that they are simply naming and referring to visuals in similar ways (and thus more keywords are found in this category). The next stage in the analysis was thus to count tagged visuals in order to determine the comparative usage. Texts included in the BAWE corpus are stripped

of tables, figures, and block quotes and a tag is left where these have been removed, making it straightforward to calculate the number of elements of each feature. In BAWE tagging, a 'table' consists of any graphic presented using rows and columns while a 'figure' covers any graph, diagram, image, picture, or drawing; both tables and figures are omitted and replaced with a tag when assignments are converted to plain text for inclusion in the corpus.

In BAWE, prose formatted as a list is tagged at the beginning and end but the list items themselves are left intact. The BAWE mark-up distinguishes between 'genuine' lists and 'false' lists; a 'genuine' or prototypical list contains a number of 'list items', each consisting of a word or noun/verb phrase and with the items separated by bullet points or similar (e.g. a hyphen), by letters or numbers, or marked as a list through the use of indentation. However, frequently assignments contain 'paragraphs of running text carrying list-like formatting' (Heuboeck et al., 2005: 29); these 'false lists' or 'list-likes' are presented as a list in the assignment, yet contain larger units of text per list item (see Figure 5.5 for example of a listlike).



Conclusions

The experiment yielded the following conclusions:

- The efficiency of a single stage centrifugal pump at high pump speed (3000 RPM) is better than it at low pump speed (2000 RPM).
- The input power with high pump speed increases faster than the one with low pump speed as discharge increases.
- The relationship between total head and discharge is not affected by pump speed, but higher pump speed provides higher total head.

Figure 5.5 Example of listlike

As the distinction between genuine lists and listlikes is 'inherently fuzzy' (Heuboeck et al, 2005: 29), for mark-up purposes, the demarcation between the two categories is viewed in terms of whether the majority of list items consist of syntactically incomplete sentences (and are thus 'lists') or complete sentences (and are 'listlikes').

Table 5.10 shows normalized counts of tagged tables, figures, block quotes, formulae, lists, and listlikes in the student corpora, and indicates that the Chinese students make greater use than the English students of four of these features.

(pmw)	tables	figures	block quotes	formulae	lists	listlikes
Chi123	751****	1216****	97	3010****	400	3,754****
Eng123	432	798	203****	1936	382	1,413

Table 5.10 Counts of visuals and lists in the two corpora (**** p<.0001)

Lists are not used significantly more in either group, and block quotes are used more in Eng123. Block quotes are not discussed further as they are relatively low in number (since there are few discursive disciplines in the corpora). The different usage of the remaining elements partially explains the lower wordcount of assignments in Chi123 (discussed in 5.2), as explanations given using formulae require fewer words than those given in prose. A great deal of information may be given succinctly in tabular format (the wordcounts used are those provided by WordSmith Tools after tables and other graphical features have been deleted). Moreover, captions to tables and figures are sometimes long, on occasion consisting of over 100 words, yet are deleted when tables and figures are tagged; this further depletes the wordcount (the use of extended captions is explored in 7.3).

Summary

This section has described the Chinese students' markedly higher use of visual features. One possible explanation for this higher usage is that employing a table, figure, list or listlike to present information in an assignment is an attractive option for Chinese students since it reduces the quantity of connected prose required. A great deal of information may be given succinctly in a table or figure, resulting in shorter wordcounts. Lists and listlikes reduce the need for connecting chunks and again reduce the wordcount. The higher use of visuals and lists thus partly accounts for the lower wordcounts of Chinese students' texts noted in 5.2. More positive explanations for the differences are that visuals and lists are viable alternative means of giving the required information, that they do so concisely, and that they perhaps

help more visual readers to process information. Since all assignments in this study are deemed by discipline lecturers to be of a proficient standard, the strategy of using tables, figures and lists appears to be, at the very least, an acceptable way of presenting information. The use of these non-prose features varies greatly depending on the discipline, and is explored further within three individual disciplines in Chapter 7.

5.4 Chapter summary

This chapter has compared the two student corpora overall in terms of the text characteristics, and through categories of keywords. The lower assignment length in Chi123 is partially accounted for by the greater use of visuals and lists since words within tables and figures are deleted from the text files along with captions to these items, and writing in lists is likely to entail the use of fewer words than writing in connected prose. The examination of keywords provides some evidence for the common assertion in the research literature that Chinese students use informal features in their writing. Informal keywords in Chi123 are limited to *lots (of), a bit of* and three informal connectors (*besides, what's more, last but not least*). A search for contracted verb forms, however, found that the English students employed these significantly more frequently than the Chinese students, indicating that the texts in Eng123 also contain evidence of informality. Few occurrences of informal items in either corpus can be accounted for by reflective sections of writing, and I propose that student academic writing may be less formal than is claimed by Biber et al. (1999) in the LGSWE. A further key category is that of connectors, and the Chinese students were found to make high use of particular connectors such as *on the other hand* and *in the long run;* this finding is thus broadly in agreement with the literature on NNS writing. While the negative key connectors *however* and *therefore* are also used in Chi123 there is a marked difference in positioning between the student groups, with these connectors usually limited to sentence-initial position in Chi123. The third key category is that of pronouns, and the Chinese students were found to make greater use of the first person plural whereas the English students tend to adopt the first person singular. A classification of concordance lines for each pronoun suggested that both groups use *I* in reflective writing; the low use of *I* by the Chinese students thus accords with the lower incidence of reflective writing in Chi123. One

finding which has not previously been reported in the literature on NNS student writing is the significantly higher use of visuals and lists by Chinese students. These features were revealed through the use of keywords and n-grams to refer to data (e.g. *the figure, according to the*), and were investigated further through counts of tagged tables, figures, and lists in the corpora. I have argued that the use of visuals and lists are viable alternative means of presenting information in undergraduate writing. Since visuals are tagged and deleted in the BAWE corpus, it is important to make use of the whole assignment PDFs in order to see how these items are used in context.

CHAPTER 6 VARIATION ACROSS YEAR GROUPS

6.1 Introduction

Chapter 5 examined the two student corpora as whole datasets. This chapter examines the same datasets divided into year groups: years 1 and 2 are grouped together for each student group, and compared with year 3. The chapter seeks to answer research question 2:

RQ 2: Are there any variations in the characteristics identified in this study between years 1/2 and year 3?

The chapter considers which of the findings revealed through Chapter 5 show variations from years 1/2, to year 3 of undergraduate study. It should be noted throughout this chapter that the data is 'quasi-longitudinal' (Granger, 2002: 11), that is, the texts from year 3 are (on the whole) not from the same writers as those in years 1 and 2 (see 4.2.3 for a discussion of the rationale behind this division). While the progress of individual students across time therefore cannot be directly stated, a comparison of year groups can suggest *tendencies* of change across year groups for Chinese and English students.

The chapter begins with a discussion of the mean assignment length, sentence length and word length, as these descriptive statistics provide an overview of the year group corpora. Section 6.3 constitutes a major section of this chapter, and is a corpus-based search of variations in use of each of the key categories from Chapter 5, namely informal items, first person pronouns, connectors, and use of visuals and lists. A study of sample pages from textbooks used in the PRC is reported on in terms of how this may influence Chinese students' use of the identified characteristics (6.4). This is followed by counts of n-gram tokens to ascertain whether there is variation in the number of different n-grams used by each student group (6.5). Finally, the chapter turns to a consideration of the 50 most frequent 4-grams found in each of the year group corpora as a means of uncovering similarity and difference in the use of lexical chunks between the two student groups. These n-grams are categorized both structurally and functionally in order to group the n-gram forms together, and thereby providing insight into the writing of the student groups.

6.2 Text characteristics

Profiling of assignments in Chi123 and Eng123 in Chapter 5 revealed that the former had a significantly lower mean assignment length (MAL). It was proposed that this was partly due to the higher use of visuals and lists in this corpus, since words within visuals are deleted from the text files, and writing in lists is likely to be more succinct than writing within connected prose. The mean sentence length (MSL) was found to be significantly higher in Eng123 yet the mean word length (MWL) was significantly higher in Chi123. This section reports on the investigation into these characteristics for the student corpora divided into year groups (Table 6.1).

Statistic	**Chi12**	**Chi3**	**Eng12**	**Eng3**
Number of assignments	89	57	436	175
Number of tokens	140,341	139,354	876,894	458,782
Mean assignment length (MAL)	1577.69	2445.31*	2011.84	2622.13*
[Standard Deviation]	[601.6]	[1321.31]	[788.46]	[1201.54]
Mean sentence length (MSL)	19.53	21.85	23.18	24.35
[Standard Deviation]	[13.7]	[14.33]	[14.68]	[14.63]
Mean word length (MWL)	4.98	4.95	4.89	4.97
[Standard Deviation]	[2.89]	[2.92]	[2.86]	[2.87]

Table 6.1 Descriptive statistics for Chi12, Chi3, Eng12 and Eng3 (significant difference between Chi12 and Chi3, and Eng12 and Eng3, using a z test, * p=.01)

For each student group, the mean length of assignments shows a significant variation between years 1/2, and year 3 (using a z-test, p=.001). A reason for this variation is that undergraduate students are required to write both longer assignments within familiar genres in the final year of study, and are also required to write within new genres such as research reports which are considerably longer than the short essays and exercises of year 1 (Nesi and Gardner, 2006). Notably, the standard deviations are very large for the year 3 MAL in each student group, indicating a wide range of text lengths; this is consistent with students producing a variety of genres. However, the MAL for Chi12 is significantly lower than for

Eng12 (using a z-test, *p*=.001), although those for Chi3 and Eng3 are not significantly different.

The MSL rises for each student group across year groups. It might be expected that all student writers produce increasingly sophisticated academic writing through, for example, the use of longer noun phrases and the inclusion of more coordination and subordination within a greater number of clauses per sentence (cf. Larsen-Freeman, 2006). The increase in MSL from Chi12 to Chi3 is significant (*p*=.01) though the increase is not significant for Eng12 to Eng3. Across student corpora, the difference between Chi12 and Eng12, and between Chi3 and Eng3 is significant in both cases (*p*=.001). It might also be predicted that each student group would use increasingly longer wordforms across year groups, as they adopt more academic language (Grant and Ginther, 2000). However, the MWL rises significantly from Eng12 to Eng3 (*p*=.001), yet falls from Chi12 to Chi3 (though not to any significant degree), that is, the Chinese students are on average using slightly shorter words in year 3 compared to years 1 and 2. Differences between Chi12 and Eng12, and between Chi3 and Eng3 are also significant (*p*=.001).The discrepancy in the MSL and MWL is intriguing, though not statistically significant (using a 2-tailed t-test with *p*=.01). A possible explanation is pursued in the rest of this section for the counter-intuitive finding that while the year 3 English texts use longer words within longer sentences than the year 1/2 English texts, the year 3 Chinese texts employ more short words and longer sentences than the year 1/2 Chinese texts.

A reason for the high MSL and lower MWL of Chi3 compared to Chi12 may be that the Chinese students are omitting smaller, grammatical words such as articles and prepositions in years 1 and 2. By year 3, as their writing gains in accuracy, these small words are more likely to be included in their writing, having the effect of both reducing the MWL and increasing the MSL. One way of identifying which words are used more frequently is through a comparison of the number of words of each letter length in Chi12 and Chi3 (Table 6.2).

	Chi12	Chi3
1-letter words	5,598	7,271****
2-letter words	24,378	24,299
3-letter words	25,005	25,566**
4-letter words	18,950	18,458
5-letter words	13,975	13,233
6-letter words	11,949	11,089
7-letter words	11,762	11,406
8-letter words	9,319	8,818

Table 6.2 Wordlengths in Chi12 and Chi3 (using log likelihood, ** p<.01; ****p<.0001)

Table 6.2 shows that for words numbering between 1 and 8 letters, the two categories in which the counts for Chi3 are significantly higher than Chi12 are in the number of 1-letter and 3-letter words (using a log likelihood test with significance levels of respectively *p* =.0001 and *p* =.01); the category of 1-letter words (which shows the greatest discrepancy) is explored below.

Four types of 1-letter 'word' are identified by WordSmith Tools: the indefinite article, first person singular, single letter forms used as abbreviations, and single digit numerals. In explaining the increase in 1-letter words in Chi3, the indefinite article can be discounted since this is used significantly more in Chi12 (counts for *a*/*an* of 3140 and 2817 respectively at *p*=.0001). Use of the first person singular pronoun, however, varies significantly from Chi12 to Chi3 and can in part account for the apparent increase in 1-letter words across year groups: per million word counts for the first person singular are 466 in Chi12 and increase to 1576 in Chi3 (see discussion of pronouns in 6.3.3). The third group of single letter items used as abbreviations also occur more frequently in Chi12 than in Chi3 (e.g. 'Where *t* is temperature in ˚*C*, *m* is moisture content in perc'). Finally, the higher use of numerals in Chi3 also has a role to play. While all numerals are by default shown in WordSmith output by the hash sign (#), they are counted per digit in the wordcount. Thus the single digit *9* is a 1-

character word, the numeral *123* is a 3-character word, and so on. The number of single digit numerals in Chi12 (i.e. *0, 1, 2, 3, 4, 5, 6, 7, 8,* 9) is 1,663 and 2,664 for Chi3, presenting a highly significant increase (p =.0001). Single digit numerals are used for purposes such as counting items ('4 experiments', 'the past 7 years'); within equations ('V = 4/3'); labelling data ('see Figure 3', 'in Table 4'); labelling physical items ('tubes 1 and 4'); and giving lists, (e.g. 'I will achieve my personal career goal by doing: 1. Find the most suitable Masters Course and University... 2. Find a part-time Marketing Assistant job ….'). The increased use of lists in Chi3 is discussed further in 6.3.4.

Summary

This section has examined the text characteristics of the corpora divided by year groups. While the MAL is greater in the year 3 texts of each student group, it is likely that this is largely due to the lengthier assignments expected of students in the final year of study. An investigation into the higher MSL yet lower MWL in Chi3 compared to Chi12 suggested that the increase may be partly due to the higher number of 1-letter 'words' counted by WordSmith. These items include the first person singular, letter abbreviations, and single digits.

6.3 Variation in the identified characteristics across the year groups

Chapter 5 identified four key categories in the keywords : the use of informal chunks, connectors, a preference for *we* over *I,* and the use of visuals and lists. The first three of these characteristics were also found in the literature on L2 writing, though the final finding of Chinese students' use of visuals and lists has not, to my knowledge, been discussed previously in the literature. This section explores these characteristics across the year groups, comparing the use of each feature in years 1/2 with use in year 3 to establish the degree of variation in the use of these features across year groups.

6.3.1 Informal items

This section traces the use of the 'informal' items discussed in Chapter 5, reporting on their use in the year group corpora. Each item classed in 5.3.1 as informal was counted for each year group corpus and the resulting figures were normalized; log likelihood tests were carried out on the raw counts of each item. In each case, items were found to be less frequently used in the year 3 corpora than the year 1/2 corpora, excepting *a (little) bit* which remained the same in Chi3 as Chi1/2 (Figure 6.1), though only two of these differences are statistically significant (*lots/a lot* in Eng1/2, *besides* in Chi1/2; Appendix E gives raw and normalized counts for all informal items).

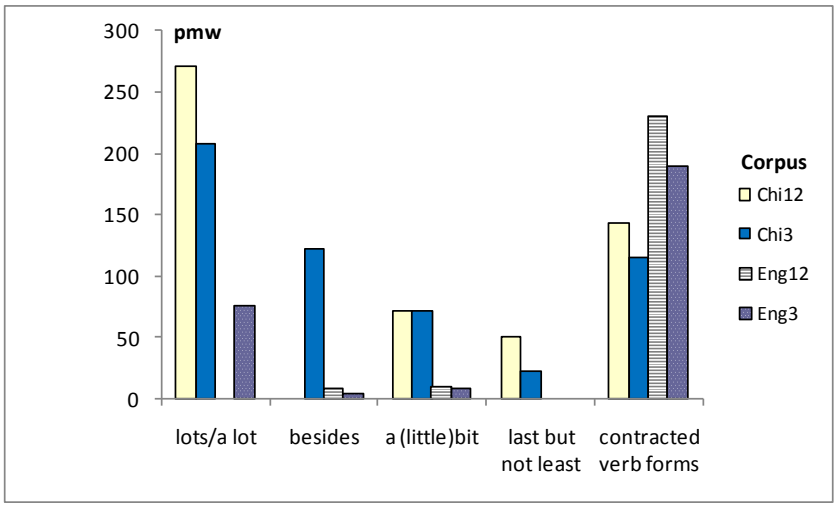


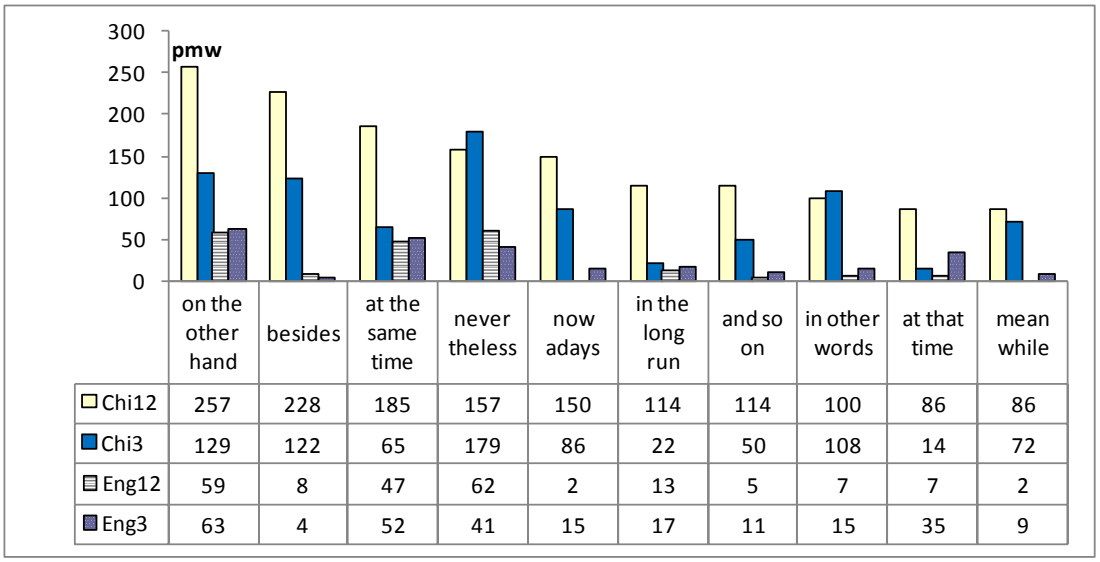
Figure 6.1 Informal items across the year groups

While few items were found in the corpora which could be classified as 'informal' language, it seems that these items are less frequently used in the year 3 corpora for each student group. The evidence here is slight, but perhaps indicates that both Chinese and English students realise these items are less commonly used in academic writing and reduce their use of them by year 3.

6.3.2 Connectors

The keywords in Chapter 5 revealed a range of single and multi-word connectors which were favoured by the Chinese students, supporting findings in the literature as to high use of fixed connecting expressions. This section reports on the investigation of the same connectors by

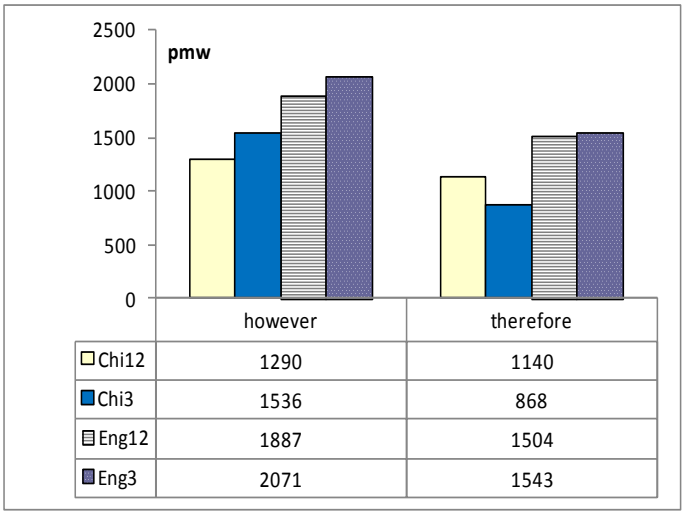
year group, finding that use of most of the connectors reduces across the year groups for the Chinese students. Connectors with a small (though not statistically significant) increase across year groups are *nevertheless* and *in other words* (Figure 6.2, results are normalized to a per million word count).



	on the other hand	besides	at the same time	never theless	now adays	in the long run	and so on	in other words	at that time	mean while
Chi12	257	228	185	157	150	114	114	100	86	86
Chi3	129	122	65	179	86	22	50	108	14	72
Eng12	59	8	47	62	2	13	5	7	7	2
Eng3	63	4	52	41	15	17	11	15	35	9

Figure 6.2 Variation in use of connectors across year groups

The negative key connectors in Chi123 (*however* and *therefore*) are shown in Figure 6.3.



	however	therefore
Chi12	1290	1140
Chi3	1536	868
Eng12	1887	1504
Eng3	2071	1543

Figure 6.3 Variation in use of *however* and *therefore* across year groups

The variation between Chi1/2 and Chi3 for the following connectors is statistically significant, and in each case the usage in Chi3 is lower than in Chi12: *on the other hand, besides, at the same time, in the long run, at that time* (in Figure 6.2), and *therefore* (in Figure 6.3). For the English corpora, *nowadays* (Figure 6.2) and *however* (Figure 6.4) were used significantly more frequently in Eng3 than in Eng1/2 (Appendix E gives raw and normalized counts for all connectors).

Unlike most of the structurally incomplete n-grams in other key categories, connectors are (relatively) fixed, semantically 'whole' formulaic sequences and therefore seem more likely to be noticed and learned as complete chunks, particularly as textbooks provide translation-equivalent sequences for each item (discussed further in 6.4). The reason for the variation across year groups may be that, for Chinese students, familiar chunks initially function as 'safe', all-purpose chunks which are applicable to a wide variety of situations (including formal and informal writing) (see discussion in 2.3). Across year groups, it is likely that students increasingly notice connecting sequences and increase their range of chunks performing similar functions. Year 3 students may have noticed that some items are informal and less commonly used in academic writing (i.e. *besides, last but not least, what's more*) and consequently are less likely to use these.

A brief case study illustrates the variation in connector usage across year groups. A small number of Chinese students (n=4) submitted work in both years 1/2 and year 3, and samples of text surrounding the use of *besides* were analyzed for one of these students in year 1 and for year 3. Examples from the first extract (268 words 12 sentences) are shown below, with connectors italicized and numbered by sentence (e.g. S1):

S1: *As regards* Oxford*,* the transport is very well-developed, convenient and well served by road …

S3: *What is more*, there are 24-hour express and frequent coaches linking Heathrow Airport…

S5: *Besides*, car park in the city centre is limited.

S8: *In addition,* the Oxfordshire county council encourage local Oxford residents to use cycle…

S10: *Additionally,* Oxford has some of the highest rates nationally for people travelling to work by…

S11: *In addition to* encouraging people to use park-and-ride services, the Oxford City Council also…

(3085a, year 1).

In this extract from year 1 writing, the student's use of connectors is limited to sentence-initial position, and the range is narrow (*in addition* is repeated and *additionally* is also used). A search for *besides* in year 3 writing by the same student found just one instance and a similar stretch of text surrounding *besides* was analyzed (275 words, 10 sentences). In this extract, the sentences adjacent to *besides* contain sentence-initial connectors, but these are more formal than those selected in year 1 and exhibit a range of functions (mainly additive in year 1, concessive and adversative in year 3):

S2: Noone and Griffin (1997) have stated *two types of* costing techniques in their study, *one is* conventional costing methods *and the other is* activity-based costing (ABC).

S2*:* *Basically*, the difference between these two costing methods is …

S3: The ABC approach does not allocate costs by using an index of volume *but* based on the activities that cause them to be incurred, and overhead costs can *then* be precisely assigned to…

S4: *Therefore*, activity-based customer profitability analysis can *then* be able to …

S6: In terms of the similarities, activity-based costing is adopted in both methods as a means to categorise costs derived from the activities rather than the product.

S7*:* *Besides that,* their focuses are both on the maximising the profit rather than the revenue.

S8: *However*, in relation to the difference, MSPA has an attempt to back up the YM decisions *while* YM is in turn…

S9: The developments of the approaches have been widely accepted *but* little practical progress has been achieved (Burgess and Bryant, 2001).

S10: The main reason is that many managers found it difficult to carry out *because of* the complexity of the data…

(3085b, year 3).

While in the year 1 extract, each sentence consisted of a single clause statement, the year 3 extract contains more complex sentences, with more cohesion within each sentence. For

example, S2 lists 'two types of…. one is… and the other is….'. The longer sentences use *but* and *then* to connect clauses.

This example illustrates a change in one student's writing. Using the figures for texts from the different year groups for the cohort overall, it seems that the group of preferred connectors in years 1/2 are less often used in the year 3 texts, and that a reason for this is an increase in the use of other connectors and of more varied ways of achieving cohesion in text.

6.3.3 First person pronouns

The third characteristic identified in Chapter 5 was the Chinese students' preference for the first person plural over the singular. A dominant use of *we* in Chi123 was in recounting methodological procedures (e.g. 'In those experiments, *we* determined the density…'). In contrast, Eng123 use the first person singular significantly more frequently than Chi123; the largest category for *I* was in reflective sections of writing (e.g. '*I* am self directed…'). This section reports on variation in the counts of *we* and *I* across the year group corpora, and also considers the functions of a 100-word sample of each pronoun in each subcorpus.

Comparison of first person pronouns in the year group corpora indicates that the year 3 Chinese students make more use of *I* and less use of *we* than the year 1/2 students (Table 6.3).

	Chi12	**Chi3**	**Eng12**	**Eng3**
we	2344	2113	1291****	630
I	466	1576****	1569****	1147
Total	2810	3689	2860	1777

Table 6.3 First person pronouns across four corpora (Statistics compare Chi12 and Chi3, and Eng12 and Eng3; **** p=.0001)

Use of first person pronouns for the Chinese students decreases across year groups, as the increase in the use of first person singular does not make up for the reduction in use of the first person plural. For English students, use of both first person pronouns decreases over the period of study. To determine the functions of the pronouns, 100 random concordance lines for *we* and for *I* were generated for each of the 4 subcorpora and then categorized as described in section 5.3.3 for the two corpora overall. This gave the results shown in Tables 6.4 and 6.5.

we	**Rep-resentative**	**Guide**	**Recounter**	**Opinion-holder**	**Reflecter**
Chi12	40	24	29	3	4
Chi3	23	37	31	5	4
Eng12	30	12	40	6	12
Eng3	49	22	17	4	8

Table 6.4 Classification of functions of *we* in four corpora (100 random lines)

I	**Rep-resentative**	**Guide**	**Recounter**	**Opinion-holder**	**Reflecter**
Chi12[31]	0	28	17	38	16
Chi3	0	12	31	8	49
Eng12	0	15	26	16	43
Eng3	0	29	29	13	29

Table 6.5 Classification of functions of *I* in four corpora (100 random lines)

Tables 6.4 and 6.5 reveal that for both student groups and also both year groups, the 'representative' function is only instantiated through *we,* whereas the 'opinion-holder' and 'reflecter' functions are mainly revealed using *I.* The 'guide' and 'recounter' functions are present in both *we* and *I* categories.

A comparison of the five functions across year groups for the Chinese students indicates that the representative category (e.g. 'There are still large areas of the bacterium's biology *we*

[31] In Chi12 there are only 57 instances of *I.* These were all classified and converted to a percentage. The total is 99% due to rounding.

don't understand') is less prevalent in the proportions within Chi3 than Chi12 (all instantiated through *we*). In contrast, the guide category (e.g. 'In this section, *we* look at…') is given through *we* rather than *I* in Chi3 (i.e. an increase from 24% in Chi12 to 37% in Chi3 for *we*, and a corresponding reduction in Chi12 from 28% to 12% in Chi3 for *I*). It seems that by year 3 the Chinese students change the way they guide the reader through the text, preferring to use the plural pronoun. The proportion of instances of *I* as opinion-holder (e.g., '*I* personally find the use of motivational skills to be…' reduces dramatically from Chi12 to Chi3 (38% to 8%). It is surprising that the Chinese year 3 texts do not contain a higher proportion of *I* to give opinions than the year 2 texts, though it should be borne in mind that the actual number of occurrences is relatively small with only 57 instances of *I* in Chi12. The use of *I* as reflecter (e.g. '*I* remember *I* could not wait…') in Chi3 at 49% is much higher than the 16% in Chi12, though there was little evidence of the kind of end reflective sections found in the English texts (i.e. sections labelled as 'reflections on the essay' or similar).

For the English students, the main difference across year groups is that the use of *we* as representative is higher in year 3 (30% in Eng12 and 49% in Eng3), the guide function becomes more prevalent in both pronouns, and the use of *I* as reflecter decreases (43% in Eng12 and 29% in Eng3). The English students' writing appears to shift from *I* as reflecter to *I* as guide, perhaps reflecting the longer genres required in year 3 (and hence the need for more explicit guidance for the reader).

However, in consideration of these variations it should be remembered that the use of *I* is significantly higher in Chi3 compared to Chi12, and the use of both pronouns is significantly lower in Eng12 compared to Eng3. One reason behind this change is undoubtedly the effect of reflective sections of writing: the reduction of occurrences of *I* in Eng3 and the reduced proportion of *I* as reflecter in Eng3 both support this.

6.3.4 Visuals and lists

Chapter 5 examined the use of non-prose features in the corpora, reporting that the Chinese students make significantly greater use of tables, figures, formulae and listlikes than the English cohort. This section reports on the use of these features in the year group corpora.

Each feature was counted in the texts for the different year groups (using the Excel contextual data sheet compiled of the corpora). This count revealed that the use of all features excepting lists is higher in Chi3 than in Chi1/2, with counts for figures and listlikes (prose written in sentences within a list structure) being significantly higher (p=.0001). (Figure 6.4).

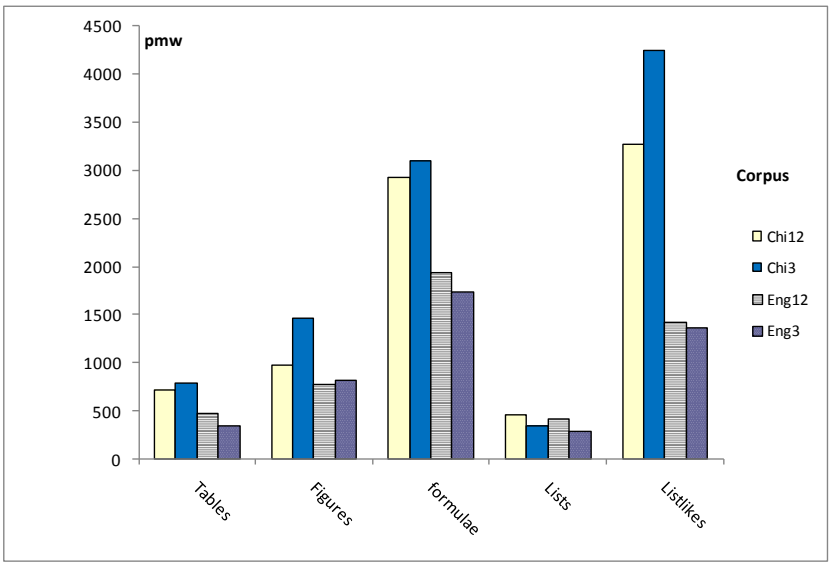


Figure 6.4 Number of visuals and lists by year group (pmw)

In contrast, the use of tables, formulae and lists are all significantly higher in Eng1/2 than in Eng3 (the count for figures shows a slight increase) (all raw and normalized counts are provided in Appendix E).

It seems that the year 3 Chinese students are increasingly likely to use bulleted or numbered items rather than paragraphs of connected prose as a presentation style in their writing. It may be that they adopt this strategy in order to meet the challenge of writing increasingly lengthy assignments in a variety of genres. If this writing strategy has proved successful in

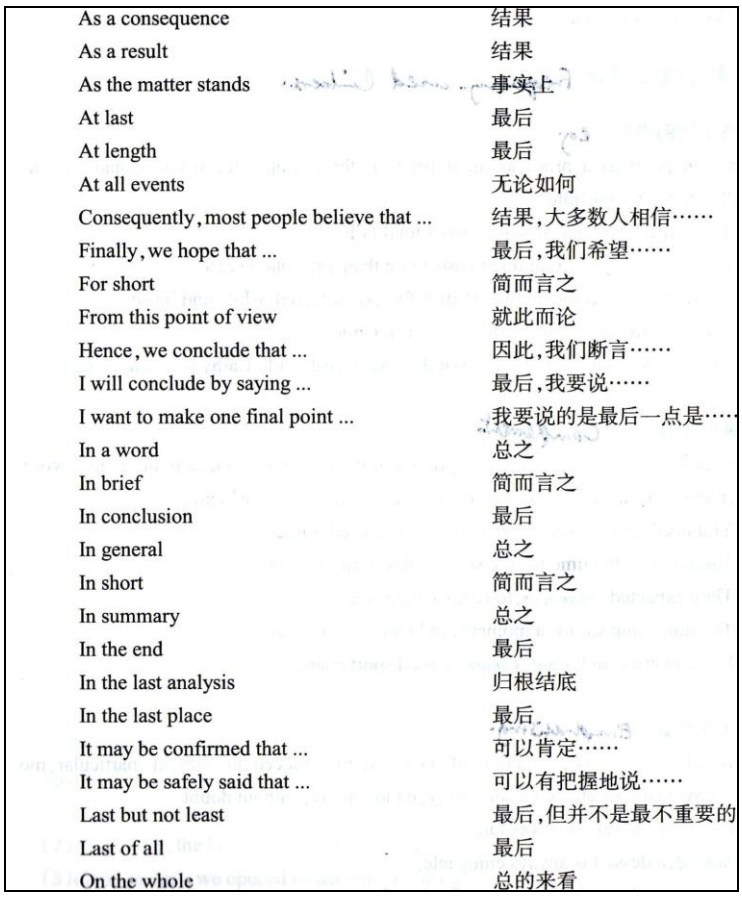
years 1/2, perhaps it is increasingly adopted in year 3. It is not possible, however, to ascertain from a corpus study whether Chinese students tend to opt for assignments in which the rubric encourages this style of writing, or whether they simply choose to answer assignments in this way. In order to understand the different ways in which the features are used in the writing, a whole text examination of texts is necessary. Chapter 7 thus examines pairs of assignments by Chinese and English students on the same disciplinary topic which employ visuals, formulae and lists differently.

6.4 Textbook study

Thus far, this chapter has suggested that use of the particular features of Chinese students' writing identified in this study (and in much of the learner corpus literature), reduces over the three years of UK undergraduate study. One reason suggested in the literature for the higher use in years 1/2 is the influence of textbooks and textbook-informed teaching (e.g. Milton and Hyland, 1999; Paquot, 2010). This influence is likely to be particularly pronounced in the PRC due to the emphasis placed on learning through textbooks and the Intensive Reading programme (as described in 1.2.3). This section reports findings from a small study of textbooks in order to hypothesize as to their priming of Chinese students' use of informal language, first person pronouns and connectors. The study comprises the whole text reading of 15 secondary school and university textbooks on English language. All textbooks are in widespread use, were published between 2002 and 2008 in the PRC, and are aimed at Chinese students preparing for the NMET (English section of the university entrance exam), College English Test or IELTS; materials are aimed at either in-class use or self study. The three scanned example pages discussed in this section are from two self study guides aimed at secondary school students preparing for versions of the NMET (each province devises its own English test). The extracts are examined in terms of the influence that the informal language, pronoun use, and connectors may have on Chinese students' later academic writing in the UK.

Example One

The first example depicts a list of connectors classed as 'endings' in English with their Chinese translation equivalents (Figure 6.5).



As a consequence	结果
As a result	结果
As the matter stands	事实上
At last	最后
At length	最后
At all events	无论如何
Consequently, most people believe that ...	结果，大多数人相信……
Finally, we hope that ...	最后，我们希望……
For short	简而言之
From this point of view	就此而论
Hence, we conclude that ...	因此，我们断言……
I will conclude by saying ...	最后，我要说……
I want to make one final point ...	我要说的是最后一点是……
In a word	总之
In brief	简而言之
In conclusion	最后
In general	总之
In short	简而言之
In summary	总之
In the end	最后
In the last analysis	归根结底
In the last place	最后
It may be confirmed that ...	可以肯定……
It may be safely said that ...	可以有把握地说……
Last but not least	最后，但并不是最不重要的
Last of all	最后
On the whole	总的来看

Figure 6.5 Example list of connectors under the heading 'conclusion'. Source: *Success with Test (*2008) Jilin Publishing Group inc.*,* p.21.

This list is a subset of an 8 ½ page list of lexical words and chunks given within functional groupings (e.g. additive, causative, comparative). No information is provided in the list as to the usage and formality of each item and the implication is that the more 'formal' connectors (e.g. *as a consequence, in conclusion*) and the less formal chunks (e.g. *in a word, last but not least*) are substitutable. Indeed, in the extract the Chinese translations for *in a word, in general, in summary* are identical[32]. Two of the connectors include the use of the first person plural (*Finally, we hope that…; Hence, we conclude that…*) and two contain the first person

32 I am grateful to Guozhi Cai for providing translations of these items and the rubric in examples two and three.

singular (*I will conclude by saying…; I want to make one final point*…), which may increase students' pronoun use in their writing.

The widespread provision of such lists is likely to contribute to the use of informal connectors by Chinese students, and to the mixing of informal and formal language generally in academic writing. Connectors, whether words or chunks, are easily-identified, semantically-whole items which are likely to be noticed by learners, particularly those who have spent time writing short test papers such as the NMET or IELTS. Such salient language is also likely to be noticed by teachers and examination markers (cf. Thewissen, forthcoming: 9, comments on raters paying 'more attention to linguistically-marked textual cohesion [e.g. connectives] than to semantic coherence').

Example Two

The second example is a test paper from a university entrance exam in Fujian province in 2005 and is provided in a writing guide for students (Figure 6.6). Students are allowed 30 – 40 minutes to complete the writing test (depending on the province). A translation of the Chinese rubric from the test paper in Figure 6.6 is provided below:

> At the moment a few students try to plagiarise in exams, and an English journal is asking for articles on this topic from secondary school students. Please write an article entitled 'My Opinion on Cheating in Examinations' based on the following prompts:

Main reasons	There are too many exams; they are too difficult.
	Students are lazy and do not work hard enough.
	Students want to please their teachers and parents.
Your opinions	Plagiarism is wrong and breaks school regulations.
	Students should be honest and work hard.
	… (other reasons)

1. The short article must include the above points, but you can include your own ideas too within reason.
2. The title and first sentence should be excluded from the wordcount.
3. You should write a minimum of 100 words.
4. (translation of 'cheat' (verb)).

附15:福建卷书面表达题(**Writing in 2005 Fujian Paper**)

目前,学校存在少数学生考试作弊现象。某英文杂志社拟对此现象向中学生征文,标题是"My Opinion on Cheating in Examinations"。请根据下列提示用英语写一篇征文稿。

内容要点如下:

主要原因	考试偏多、偏难
	不用功、懒惰
	取悦父母、老师
个人看法	作弊不对、违反校规
	要诚实、努力学习
	……(其他看法)

注意:1. 短文必须包括所有内容要点,可适当发挥;

2. 短文标题与开头已为你写好,不计入词总词数;

3. 词数:100左右。

4. 参考词汇:作弊 cheat (*v.*)

My Opinion on Cheating in Examinations

It is known to us all that some students cheat in examinations at school.

...

【参考范文(*One possible version*)】

My Opinion on Cheating in Examinations

It is known to us all that some students cheat in examinations at school.

As students, we often take examinations at school, but sometimes we have too many examinations which are too difficult for us. On the other hand, some of us are lazy and don't work hard at their lessons. So when taking examinations, they sometimes cheat in order to get better results to please their parents and teachers.

In my opinion, it is wrong to cheat in examinations because it breaks the rules of schools. We students should be honest and try to get good results by studying hard instead of cheating in examinations. What's more, we should improve our study methods and get well prepared for examinations.

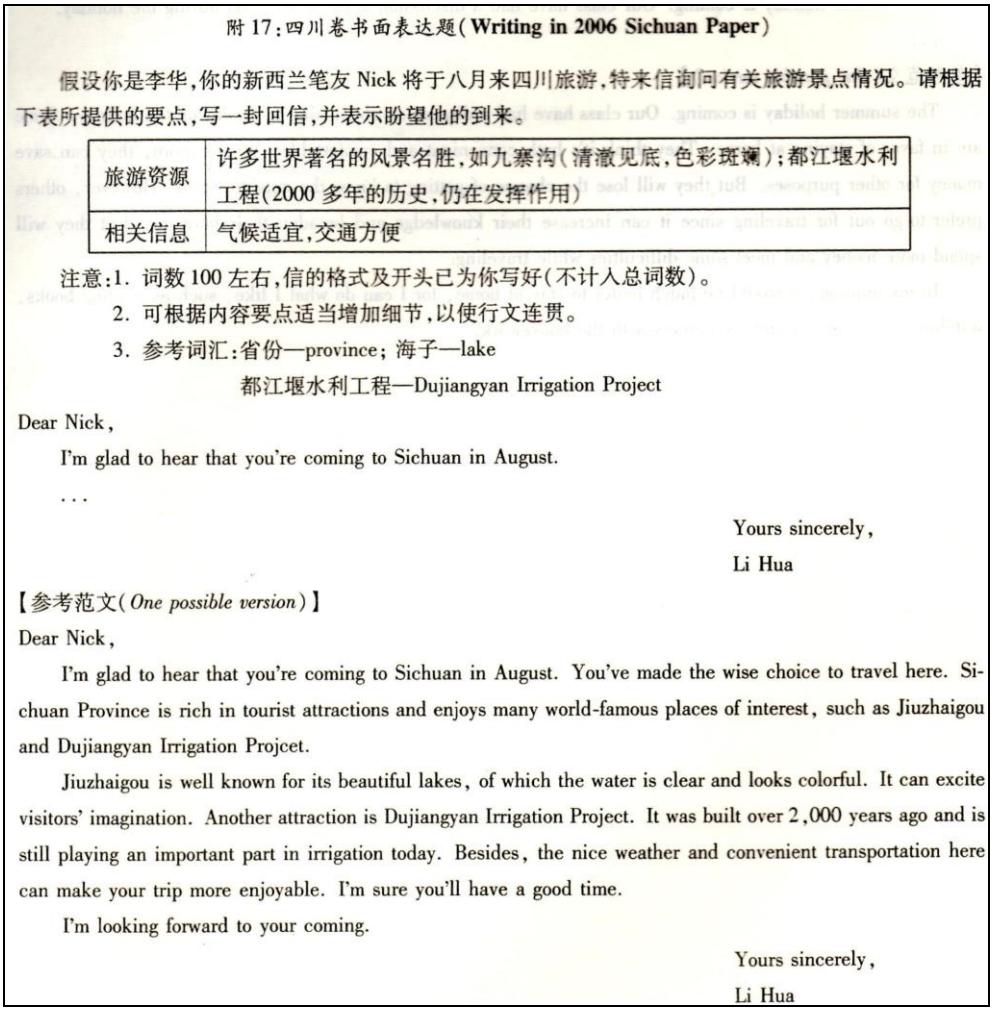
Figure 6.6 Example Two: Question from Fukian NMET, 2005 with sample answer, scanned from: *A Guide to English Writing for Senior High School Learners,* 2006, Hubei Education Press p.198

The question title illustrates the general knowledge nature of the writing test and the structured layout of the model essay is typical of practice test books. The rubric provides not only guidance as to the *type* and *structure* of the required answer, but also extensive prompts, rendering the student's task one primarily of translation. The model answer includes each point from the table of 'prompts', in the same order, and includes the provided 'opinions'. The sample 120-word answer illustrates extensive use of inclusive *we* as in *'we* students', '*we* often take', '*we* should improve'. Interestingly, pronoun use alters to the third person plural when the writer considers students who cheat. Thus, although 'we' all have exams and may experience difficulties, it is a subset of this collective group ('some of us') who are 'lazy' and the reader is informed that '*they* sometimes cheat': the implication is that the reader is not a member of the subset of cheaters.

This short sample answer includes two connectors which were found statistically more frequently in Chi123 (*on the other hand, what's more)*, and it is argued that students are encouraged to make use of these chunks through the inclusion of such connectors and the first person plural in a model answer for a high stakes examination paper.

Example Three

As well as short argumentative essays, a common format for the NMET is a short letter. Example Three gives similarly detailed instructions as to layout and content (Figure 6.7).



附 17：四川卷书面表达题（**Writing in 2006 Sichuan Paper**）

假设你是李华，你的新西兰笔友 Nick 将于八月来四川旅游，特来信询问有关旅游景点情况。请根据下表所提供的要点，写一封回信，并表示盼望他的到来。

旅游资源	许多世界著名的风景名胜，如九寨沟（清澈见底，色彩斑斓）；都江堰水利工程（2000 多年的历史，仍在发挥作用）
相关信息	气候适宜，交通方便

注意：1. 词数 100 左右，信的格式及开头已为你写好（不计入总词数）。

2. 可根据内容要点适当增加细节，以使行文连贯。

3. 参考词汇：省份—province；海子—lake

都江堰水利工程—Dujiangyan Irrigation Project

Dear Nick,

I'm glad to hear that you're coming to Sichuan in August.

...

Yours sincerely,

Li Hua

【参考范文（*One possible version*）】

Dear Nick,

I'm glad to hear that you're coming to Sichuan in August. You've made the wise choice to travel here. Sichuan Province is rich in tourist attractions and enjoys many world-famous places of interest, such as Jiuzhaigou and Dujiangyan Irrigation Projcet.

Jiuzhaigou is well known for its beautiful lakes, of which the water is clear and looks colorful. It can excite visitors' imagination. Another attraction is Dujiangyan Irrigation Project. It was built over 2,000 years ago and is still playing an important part in irrigation today. Besides, the nice weather and convenient transportation here can make your trip more enjoyable. I'm sure you'll have a good time.

I'm looking forward to your coming.

Yours sincerely,

Li Hua

Figure 6.7 Example Three: Question from Sichuan NMET, 2006 with sample answer, source as Example Two, p.211

A translation of the Chinese rubric in Figure 6.7 is provided below:

Imagine that your name is Li Huan and you have a penfriend from New Zealand called Nick. Nick will come to Sichuan province in August. He has sent a letter to you asking about tourist attractions. Based on the following prompts, write a letter back to Nick:

Tourist attractions	Many world-famous places of interest e.g. Jiuzhaigou, (where the water is crystal-clear and splendid) and Dujiangyan Irrigation Project (2,000 years old and still in use).
Other relevant information	Fine weather, convenient transport links.

1. Write a minimum of 100 words. Use the prompts provided.
2. The format of the letter and first sentence are provided. Exclude these from the wordcount.
3. You should write a minimum of 100 words.
4. (translation of 'province', 'lake', and 'Dujiangyan Irrigation Project').

The model letter makes use of pronouns within contracted verbs (*I'm, you've, I'm sure, you'll*) and the connector *besides* (a keyword discussed in 5.4). Despite the presence of contracted verb forms in this (typical) example text, my earlier analysis revealed that Chinese students rarely use these in their writing; this lack of use may be due to the verb forms only occurring within example letters and not in the short essays. As with Example Two, the student's role is limited to translating the provided information and making these items cohesive, and there is little scope for originality. The writing task in NMET appears to be one of translation rather than the production of writing in the student's own words. A study by Wang and Wen (2002) lends support to the importance of translation. Their research featured 16 Chinese EFL learners in a think-aloud study, and found that the learners used L1 an average of 30% of the time in their think-aloud. Wang and Wen argue that 'the L2 writing process is a bilingual event: L2 writers have two languages… at their disposal when they are composing in L2' (p.239). In terms of lexical priming, this means that switching back and forth between languages, students are constantly aware of the primings of similar words and chunks in their L1.

While students continuing with their studies in the UK will recognize that writing a letter is a different task to writing an essay, the extensive letter-writing practice carried out for the NMET may well affect students' later language use in terms of the employment of informal language, pronouns and, connectors.

Summary

The three example pages illustrate the features of Chinese students' writing in English which have been highlighted in both this study and in the available literature, namely, use of informal language, *we,* and particular connectors. Since short essays and letters are presented as model answers to high stakes examination questions, students are likely to be primed to believe these chunks are acceptable in academic writing and to use them in their longer undergraduate assignments in the UK. This blurring of formality levels and high use of particular chunks is especially likely at the start of their UK academic career when students perhaps lack the breadth in academic reading required to distinguish between genres. By year 3, students are likely to have received feedback as to the informal nature of aspects of their writing, to have read and noticed which pronouns are most common in academic writing, and to have broadened their range of connectors through increased exposure to academic writing (cf. Hoey's, 2005, notion of the lower range of primings encountered by NNSs, discussed in 3.2).

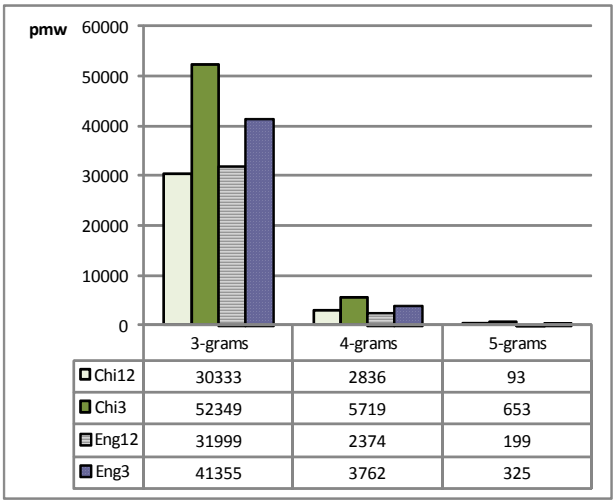
The next two sections return to a consideration of n-grams in the year group datasets.

6.5 N-gram tokens

This section compares the variation in the numbers of 3, 4, and 5-grams occurring at least 20 times pmw and in 5% of assignments (see discussion of these frequency and dispersion thresholds in 4.3.3). Over time, it would be expected that the number of n-gram tokens may rise as students add to their repertoire of frequent chunks, rather than relying on a few similar chunks. While the year group corpora are quasi-longitudinal since the student groups in each cohort are different, the token counts in Figure 6.8 indicate that there are more n-grams in year 3 than years 1/2 for both student groups.

In terms of the comparison across student groups, it appears that while the figures for the year 1/2 corpora are similar, the token counts for Chi3 are higher than for Eng3. These findings should, however, be interpreted with caution. The higher wordcounts of the year 3 corpora mean that it is more likely that n-grams in these datasets will meet the dispersion

threshold of 5% of assignments. In particular, the fact that there are 89 assignments in Chi12 and 57 assignments in Chi3 (yet the same number of tokens) affects the calculation. The parameters of 20 pmw and 5% of assignments in Chi12 require three n-grams in five different assignments to reach this threshold (effectively raising the number of n-grams to 5), whereas in Chi3 the same parameter equates to 3 n-grams in just three different assignments.



	3-grams	4-grams	5-grams
Chi12	30333	2836	93
Chi3	52349	5719	653
Eng12	31999	2374	199
Eng3	41355	3762	325

Figure 6.8 Comparison of n-gram tokens in the four corpora

Omitted/substituted functional words

A count of n-grams may also be misleading as only those n-grams with the same outward form are counted. For example, it may also be the case that on occasion functional words are missed out of longer n-grams, particularly in the Chinese corpora, meaning that the n-gram counts are distorted as some 3-grams could be viewed as 4-grams with a missing element.

However, unless an individual repeats these chunks, or several individuals produce the same 'erroneous' chunks, then these will be hidden in a frequency-based extraction method. Chunks with omitted elements were not found in sufficient numbers to account for the MWL discrepancy described in 5.2, though they are interesting in analyzing variation within individual students' use of chunks. To determine the frequency of what Wiktorsson terms

'erroneous target language prefabs' (2003: 158), a search was conducted for words from the most frequent ten 4-grams in Chi123, then left or right sorted to find curtailed n-grams. For example, searching for *end* in Chi3 gave the following concordance line with a missing article:

(1) However, this is not *the end of* story. A term of 'collapsing bubble' needs to be introduced to equation… (0279a).

This would not feature in the figures for the 4-gram *the end of the* as the final definite article is missing. Similarly, the following example is omitted from the counts for *on the other hand*:

(2) …because substrate was used continuously and would run out eventually. *On another hand,* increasing amounts of phosphate produced would inhibit the enzyme and slow… (0100a).

Similarly, the following examples (from the same student) show an omission of the article, resulting in a 3-gram which does not show up in the 4-grams for the chunk *in the long run*:

(3) …the population with no dispersion is likely shrinking in population size *in long run*. The ones with 0.5 and 0.8 dispersal rate maintained slightly above the (0036c).

(4) …event rates might result in a conspicuous mean of population sizes *in long run,* but with a comparatively low frequency. (0036c).

While still a 4-gram, the following example illustrates the issue of variations of 4-grams not showing up in the frequency results. There are ten occurrences of *last but not least* in Chi123 and a single instance with *lastly* instead of *last*:

(5) *Lastly but not least,* the biggest problem of guideline is its inflexibility. (7001c).

Using an indefinite article rather than the definite article also results in 'missing 4-grams', assuming that this does not occur often enough to form a 4-gram itself. The example below contains two such chunks:

(6) *In a short-run* this shift appeared to be an effective way to boost up economy, since fiscal policies take little time to take effect; but *in a long term*, excessive reliance in the public sector not only caused heavy government deficit, but also... (7034a).

A search for *time*, sorted to find instances of *same* revealed different preposition usages; these reduce the 4-gram counts for *at the same time*:

(7) *In the same time*, we should pay attention on the samples per frame, number of bits per… (6107b).

(8) The traceback starts at a given state from irrespective of state metrics. *In the same time* we noticed that the model also runs faster than previously… (6107b).

(9) …information (URL), ASDA aims to take a fair and objective method, *on the same time* ensure a minimum acceptable standard of performance is achieved. (3018c).

The examples here show missing words and substituted grammatical words from frequent 4-grams. The discussion in this section assumes that the student intended to use the more common 4-gram (note no instances of these kinds were found in the English students' assignments). The use of 3-grams which are more commonly 4-grams such as *on other hand*, or 4-grams with nonstandard prepositions such as *on the same time* may be due to the Chinese students' internalizing the chunks in this way; that is, it becomes a chunk for the student. Alternatively, perhaps a shorter chunk such as *other hand* or *same time* has been remembered with the additional prepositions and articles built in at the time of writing.

6.6 Classification of four-grams

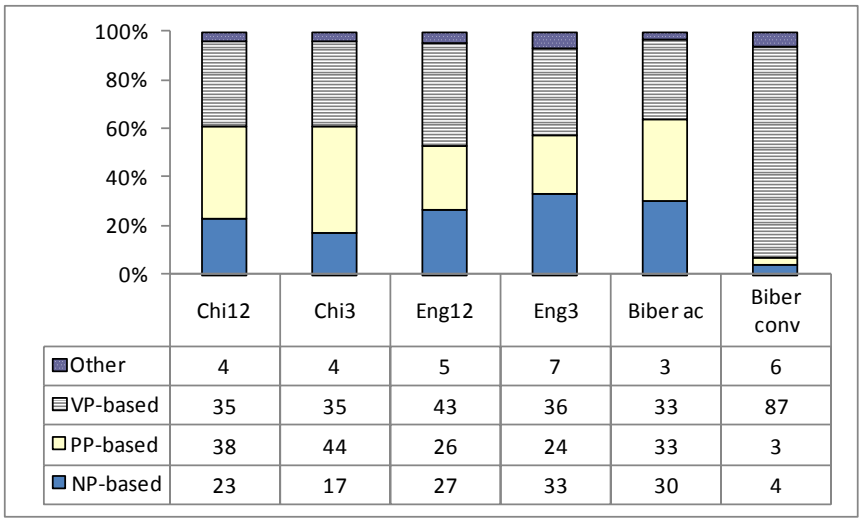
This section details the extraction and classification of frequent 4-grams in the two student corpora as a means of exploring similarity as well as difference within the datasets. Four-grams were chosen for analysis as they are frequent enough to yield sufficient examples, and long enough to facilitate classification (cf. Chen, 2009; Cortes, 2004, 2006; Hyland, 2008,a,b. The most frequent 50 four-grams were first extracted from each of the four year group corpora using the thresholds of 20 occurrences per million words and dispersion across at least 5% of the texts, and from a minimum of three individuals, and checked for any overlapping 4-grams (discussed in 4.3.4). The resulting 4-grams were then classified both structurally and functionally.

6.6.1 Structural classification

The most frequent 50 four-grams in each of the four year group corpora were classified according to their structural patterns, following the description in Biber et al, (1999). Chen and Baker (2010) adapted Biber et al.'s classification system, grouping patterns together as

NP-based, PP-based, or VP-based in order to comment on the 'verb-based' nature of student writing. The classification followed here uses both Biber et al.'s academic writing patterns and Chen and Baker's broad categories (as given in Table 3.1 in 5.3.1). In this section I first consider Chen and Baker's broad structural categories, and then examine the more detailed categories from Biber et al.

The broad structural classification is given for the four year group corpora, and also Biber et al.'s categorization of professional academic writing and conversation, in order to provide some perspective in considering the nature of student writing. However, Figure 6.9 reveals that both the Chinese year groups and Eng3 have very similar proportions of VP-based chunks to Biber et al.'s academic corpus; that is, their writing is not significantly more 'verb-based' than the writing of professional academics. In contrast, Eng12 has a higher proportion of VP-based chunks (43%) than the academic corpus (33%), though this is nowhere near the 87% VP-based categorization of Biber et al.'s conversation category.



	Chi12	Chi3	Eng12	Eng3	Biber ac	Biber conv
Other	4	4	5	7	3	6
VP-based	35	35	43	36	33	87
PP-based	38	44	26	24	33	3
NP-based	23	17	27	33	30	4

Figure 6.9 Structural categorization using broad VP, PP, NP groupings

To determine the reason for the discrepancy in usage within Eng12, I then considered the more detailed categorization following Biber et al. (Table 6.6) (the structural classification for each n-gram type is given in Appendix F).

Broad category	Structural pattern	Chi12	Chi3	Eng12	Eng3
NP-based	(1) NP + *of*-phrase fragment	17	14	16	23
	(2) NP + other post-modifier fragment	4	3	7	7
	(3) pronoun/NP (+aux) + be	2	0	4	3
PP-based	(4) PP + embedded of-phrase fragment	9	18	19	19
	(5) other PP fragment	29	26	7	5
VP-based	(6) anticipatory *it* + VP/AdjP (+complement clause)	7	13	12	14
	(7) passive verb + PP fragment	5	5	10	11
	(8) be + NP/AdjP	16	6	1	3
	(9) (NP+) (verb +) *that*-clause fragment	2	3	8	1
	(10) (V/Adj+) *to*-clause fragment	5	8	12	7
Other	(11) other expressions	4	4	5	7

Table 6.6 Structural categorization of chunks (as % of most frequent 50 4-gram tokens)
Key: AdjP = adjective phrase, NP = noun phrase
P*P*= prepositional phrase, VP = verb phrase

Table 6.6 reveals that the high use of VP-based n-grams in Eng12 (and to a lesser extent in Eng3) is in part due to the greater use of category (9) ([NP+] [verb+] *that*-clause fragment), in n-grams such as *can be seen that* (61 raw occurrences in Eng12), *that there is a* (41 occurrences), *that there was a* (21 occurrences)*:*

(10) Comparing the individual results for this against the class results, *it can be seen that* a very low colony count was obtained (6004f).

(11) Looking at the results *it can be seen that* both methods gave a fairly similar count as to the number of E.coli organisms present… (6085d).

(12) It can also be shown *that there is a* limit to the longitudinal force on a tyre of… (0228h).

(13) The dot plot in figure three shows *that there is a* fairly even spread of all species of ants over different… (6035a).

(14) It is likely *that there was a* low literacy over this part of the population… (0414a).

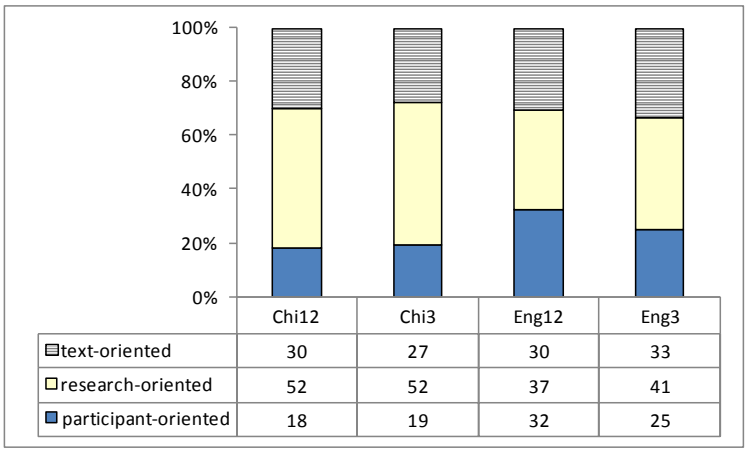
(15) He also claimed *that there was a* breach of natural justice in the granting of the planning permission (0191d).

While these categories are broadly 'verb-based', the majority of instances are from passive constructions (as illustrated by examples 10-12) and few of the verb-based n-grams are used with a personal pronoun. These VP-based 4-grams are thus dissimilar to the conversational chunks extrapolated by Biber et al. (e.g. *I don't know why, can I have a*).

The most striking difference between the two student corpora is in category (5) 'other PP fragment'. In Chi123, this category consists of two types of n-grams: connectors such as *on the other hand, in the long run,* and *in order to* + verb (e.g. *be/find/avoid*). Chinese students' use of particular connectors has been discussed in 5.3.2 and 6.3.2. The difference in use of the 3-gram *in order to* between the two corpora overall is statistically significant (194 raw occurrences in Chi123, 653 in Eng123, p=.0001). Moreover, *in order to* is used significantly more often in Chi12 than in Chi3 (116 and 78 occurrences respectively, p=.01).

6.6.2 Functional classification

The second means of classifying the 4-grams is through their metafunctions (Halliday, 1994). Figure 6.10 gives my functional classification of 4-grams using Hyland's (2008a,b) three-way system, which draws on Biber et al. (1999) and Halliday (1994) (described in 3.5.2). Figure 6.10 again shows the categorization as a proportion of the tokens for each year group corpus.



	Chi12	Chi3	Eng12	Eng3
text-oriented	30	27	30	33
research-oriented	52	52	37	41
participant-oriented	18	19	32	25

Figure 6.10 Broad functional categorization of chunks

Between the student groups, the greatest difference is in the higher proportion of research-oriented n-grams and lower proportion of participant-oriented n-grams in Chi12 and Chi3. Across the year groups, Figure 6.10 reveals that the Chinese group varies very little from years 1/2 to year 3 in terms of the use of broad metafunctions. The English group slightly increase the proportion of research-oriented and text-oriented chunks and slightly reduce the number of participant-oriented chunks. Hyland's findings (2008b) on the proportions of chunks in each category showed a much clearer variation across groups for his corpora of Hong Kong masters and doctoral students' writing, and writing from professional academics (of unknown L1 and in a range of journals), all writing in English. If it is assumed that there is an upwards gradient in English language proficiency level from masters theses to PhD dissertations, to professional research articles, then Hyland's study suggests that the proportion of 4-grams in the research-oriented category decreases and the other two categories increase as writers become more proficient in academic writing.

Each metafunction was also divided into subcategories to see whether there were greater differences at this level in the corpora. Although all instances of a 4-gram may not have the same function, for the purposes of this investigation each 4-gram type has been allocated to a single functional category (see 3.5.2 and 4.3.4 for discussion on difficulties with a monofunctional classification). The resulting categorization is given in Table 6.7 (the functional classification for each four-gram type is given in Appendix F).

The more detailed categories are discussed under the broad category headings.

Participant-oriented

While the total proportion of participant-oriented n-grams barely alters (18% in Chi12, 19% in Chi3), the proportion within engagement and stance categories reverses. In Chi12 the majority of the participant-oriented n-grams relate to the writer's stance, for example *(is due) to the fact that, it is believed that* and relatively few relate to the engagement of the reader in the writing. In Chi3, more engagement n-grams are used, for example *it can be seen (that), it is possible to, it is important to.*

		Chi12	Chi3	Eng12	Eng3
participant-oriented	engagement	6	13	13	13
	stance	12	6	19	12
research-oriented	description	23	14	13	20
	location	9	11	8	7
	procedure	8	16	9	10
	quantification	9	10	7	4
	topic	3	1	0	0
text-oriented	framing	10	6	13	12
	resultative	4	6	6	5
	structuring	0	3	4	6
	transition	16	12	7	11

Table 6.7 Functional categorization of n-grams
(as % of most frequent 50 four-gram tokens)

Research-oriented

For all four corpora, the research-oriented chunks are mainly in Hyland's (2008a,b) research-description category, followed by the other subcategories in approximately equal proportions. Students describe elements of their work (e.g. *the nature of the, the form of the)* and quantify elements of the work (e.g. *the rest of the, one of the most)*. The research-descriptive category accounts for the higher proportion of n-grams in the research category for the Chinese datasets, as students describe a study they have carried out. A common group of chunks in Chi123 are those containing *in order to* + verb, and these were categorized as research-procedural (discussed earlier in the consideration of structural categories).

Text-oriented

The text-oriented chunks in this section show differences between the student groups. In Chi123 the largest subcategory within the metafunction of text-oriented is transition, followed by framing, though the distinction becomes less apparent across year groups. Transition signals in Chi123 are fixed, semantically whole chunks such as *in the long run, last but not least* and *on the other hand* in Chi12 and *on the other hand* and *in the short term* in Chi3.

The only semantically complete chunk in this category in Eng123 is *on the other hand*, and this has a (proportionally) lower frequency count than for the Chinese corpora.

Summary

The categorization of 4-grams has indicated that neither group of student writing includes a high proportion of 'verbal' chunks. Most n-grams in the VP-based category are from passive constructions, and are unlikely to be found within conversation. The analysis of 4-grams confirms previously-discussed differences such as the Chinese students' greater use of particular connectors and also indicates that *in order to* + VP is a favoured sequence.

6.7 Chapter summary

This chapter has reported on the investigation of aspects of the writing in terms of variation across the year groups. The chapter first examined text characteristics, suggesting that the longer MAL in year 3 texts is due to the lengthier genres of assignments required. The longer MSL and MWL in Chi3 compared to Chi12 is in part due to the greater use of numerals in year 3. The chapter then focused on the examination of the key categories from Chapter 5: informal language, connectors, first person pronouns, and visuals and lists. While the year group data is quasi-longitudinal, some tendencies for variation across year groups can be given, and it seems that the Chinese students reduce their use of most of the distinguishing features across year groups. The use of informal chunks, particular connectors, and the first person plural was found to decrease from years 1/ 2, to year 3 in the Chinese student group. Conversely, the use of visuals and lists in assignments increased over the undergraduate years. The inclusion of 4-gram classification in this chapter gave a further means of interrogating the data, and also provided a way of investigating similarities in the student corpora, as well as the differences uncovered through keyword analyses. This analysis indicated that the student corpora have a similar number of verb-based 4-grams to a professional academic corpus, and that the Chinese students make greater use of research-oriented sequences such as *in order to.*

Throughout the analysis in this chapter, differences in the composition of the student corpora have became increasingly apparent in terms of different numbers of individual students in each corpus, and the proportions of assignments within disciplines. Differences of this kind are unavoidable in corpora of naturally-occurring language where the whole text is included rather than only same-size extracts, and particularly where the data is limited in some way (here by the number of assignments available from Chinese students). One way of countering these limitations is by dividing the corpora in a different manner and comparing the two student groups' writing across disciplines rather than across year groups: the next chapter investigates the effect of disciplinarity on the characteristics identified in the corpora.

CHAPTER 7 DISCIPLINARY INFLUENCES

7.1 Introduction

Recent research in student writing has emphasized the degree to which university students have to meet the challenge of writing in conventionalized ways within their disciplinary areas in order to achieve success (Bazerman, 2001; Harwood and Hadley, 2004; Hewings, 1999; Hyland, 2008b; Lillis, 1999, 2001; North, 2005b; Prior, 1998; Rai, 2008). Despite this recognition of its importance, disciplinarity within undergraduate student writing has been relatively little explored, though some studies have been conducted since the compilation of the BAWE corpus (e.g. Bruce, 2010; Gardner, 2008; Thompson, 2009). The literature on NNS undergraduate writing within different disciplines is even narrower, as most studies either consider all texts together, regardless of discipline (e.g. Chen's, 2009, study of Chinese assignments in BAWE), or are conducted on a single discipline (e.g. Lee and Chen's, 2009, study of writing in Linguistics). This chapter reports on an investigation of a subset of the assignments in Chi123 and Eng123, considering texts from three of the disciplines: Biology, Economics and Engineering. These disciplines were selected as they each contain relatively high numbers of texts by Chinese students from across the three year groups, and are not dominated by texts from a few individuals. The three disciplines are towards the 'hard' end of the 'soft-hard' dimension, as outlined in 1.3 and 2.5 (and visually represented in Figure 4.1), reflecting the discipline choices of Chinese students in UK universities for more practical areas of study. One reason for Chinese students favouring 'hard' disciplines when studying internationally may be that language production plays a lesser role than in the 'soft' disciplines of, say, History or Philosophy (see discussion of this point in 4.2), and this is supported by the prevalence for visuals and lists in these disciplines (discussed in 7.2.5 and 7.3).

The chapter examines writing within student groups across the disciplines in order to respond to research question 3:

RQ 3: In what ways do disciplines affect the identified characteristics of Chinese undergraduate writing in English?

In answering this question, the three selected disciplines are first compared to all other undergraduate disciplines in Chi123 and Eng123 using the keyword procedure to establish disciplinary differences, regardless of the L1 of students (7.2). Following this, the key categories from Chapter 5 are revisited to determine whether there are differences in these across the student groups for each discipline. Corpus searches of items from Chapter 5 are carried out to determine areas of difference; this is supported by a keyword comparison of student groups within each discipline. No evidence of different use of informal language was found either within the disciplines or across student groups within each discipline, and this category is not discussed further. The comparison of discipline subcorpora is thus limited to the use of connectors, first person pronouns, visuals and lists. In the final section, visuals and lists are compared within whole texts as this offers insights into multimodal aspects of the data (7.3).

7.2 Keyword analysis

This section outlines the wordcounts in the three disciplines, and then examines the keywords.

7.2.1 The data

For the investigation in this chapter, texts from Chinese students studying Biology, Economics and Engineering were extracted from Chi123, and reference corpora were compiled from the same disciplines within Eng123. Full text and word counts are given in Table 7.1 for each discipline. The subcorpora in this chapter are termed Chi-Biology, Eng-Biology, and so on.

	Chinese			English		
Discipline	**No. texts**	**No. tokens**	**Mean length**	**No. texts**	**No. tokens**	**Mean length**
Biology	18	33,633	1868	83	173,412	2089
Economics	20	38,086	1904	22	52,158	2371
Engineering	20	35,627	1781	97	203,782	2101

Table 7.1 Texts and wordcounts in each discipline subcorpus

No significant differences were found in the mean assignment lengths of the disciplines across either disciplines or student groups, and this statistic is not considered further.

7.2.2 Student writing in Biology, Economics and Engineering

In this section, keyword analysis is used to find words and n-grams which are significant in each of the three disciplines compared to the rest of the student corpora used in this study (Chi123 and Eng123) (Table 7.2).

Biology	**Economics**	**Engineering**
#, were.	*price, demand, monopoly, we, than, capital, increase, higher, exchange, inflation, labour, economic, unemployment, prices, countries, money, production, cost, investment, interest, firm, foreign, crisis, trade, wage, long, marginal, F, country, Y, run, elasticity, domestic, variables, currency, goods, exam, equilibrium, expectations, rates, short, consumers, monetary, surplus, policies, consumer, efficiency, spending, scale, fiscal, productivity, Phillips, slope, bank, central, monopolist, saving, relative.*	*stylus, disc, cantilever, #, amplifier, gauges, drag, moment, mechanical, shaft, gauge, deflection, modelling, discharge, motor, loading, cylinder, capacitor, measurement, bottom, sensor, angle, hole, bridge, analogue, experiment, carbon, circuit, temperature, measured, measuring, efficiency, pressure, length, values, display, digital, resistance, input, head, centre, heat, signal, air, using.*

Table 7.2 Keywords in three disciplines

For each discipline, tokens were found which are key for *both* student groups when compared to all undergraduate writing in the study (minus the discipline in question) yet are *not* key when the student groups are compared to each other (see 4.3.2 for theoretical discussion of keywords and full details of this procedure). The resulting keywords are given

in descending order of keyness for each discipline. Categories from the set of keywords for each discipline are discussed below:-

Biology

Just two keywords were found which fulfilled the criteria of being key in both Chi-Biology and Eng-Biology when each was compared to the whole corpus (minus Biology). The first keyword indicates that numbers are key in this discipline (these are also key in Engineering). Numbers in Biology are used to give quantities of data, to provide ratios or percentages, to label visuals, and so on, for example:

(1) *36* curly winged and *5* straight winged flies were counted by us. That is a ratio of *36:5* or *7.2:1* (0035a).
(2) When L=*3.0,* the effect of a *20%* reduction in k was a *10%* reduction in growth rate (table *1*) (6214c).

The second keyword for Biology (*were*) suggests that possibly the past tense and/or passive voice may be used more frequently than in other disciplines. A list of the top ten first left and first right collocates in Biology (from Chi-Biol and Eng-Biol combined), indicates that *were* is mainly used in passive constructions to recount the procedures carried out (Table 7.3).

Rank	L1	Centre	R1
1	THEY	WERE	FOUND
2	THERE		NOT
3	PLANTS		USED
4	THAT		ADDED
5	SAMPLES		RECORDED
6	CELLS		OBSERVED
7	RESULTS		IN
8	BACTERIA		CARRIED
9	FLIES		ALSO
10	AND		CAUGHT

Table 7.3 Collocates of *were* in Biology

The prevalence of *were* in recounting procedures is confirmed by a plot dispersion showing bursts of the keyword in methodology sections of texts as students provide detailed descriptions of what they did. For example in the following extract, the student describes the process of growing plants in order to extract plant material:

(3) Impatiens noli-tangere seeds […] *were sown* in John Innes II Potting Compost at the [university name] in September 2005. The plants *were grown* under approximately 50% shade in 9cm pots and re-potted to 15cm pots on 12 June 2006. Plant material, leaves, flower buds and immature pods *were collected* from 26 plants …. (6215f).

The comparison of first person pronouns in 7.2.4 below reveals that Biology uses these the least of the three disciplines, providing further support for the relatively high use of the passive voice.

Economics

In Economics, categories of keywords include aspects particular to the discipline such as jobs (*labour, unemployment, firm*); consumer spending (*demand, goods, consumer(s), spending*); and making comparisons (*than, increase, higher, relative*). The latter group are often used to describe fiscal change, for example:

(4) …then the UIP will capture this and domestic interest rate is *higher than* foreign interest rate (0076b).
(5) However, an *increase* in wages, leads to an *increase* in prices and hence an *increase* in inflation (0399b).
(6) Firms can adjust prices *relative* to each other through these different pricing strategies… (7020b).

A few keywords in Economics are from longer connectors: *long, run, short* are key due to their presence in the n-grams *in the short run, in the long run, in the short term, in the long term. In the long run* is the only connector from those discussed in 5.3.2 to be used significantly more frequently in one discipline. *In the long run* (and the related *in the short run, in the short term*) occur significantly more frequently in both Chi-Economics and Eng-Economics than in either Biology or Engineering (p=.0001). For both student groups, these sequences are most frequently found in Economics writing in discussions of future trends, lending support to Bloor and Bloor's (2001) suggestion that this discipline is concerned with prospecting the future. For example:

(7) *…in the long run,* changes in aggregate demand will have a smaller or even no effect on output and employment… (0298b).
(8) .. economic variables such as output and unemployment *in the short term* deviate from their long-term natural and potential rates…. (0202i).

The pronoun *we* is key for both student groups in Economics (Table 7.4).

Rank	n-gram	Freq.
1	WE CAN SEE	15
2	WE FIND THAT	11
3	WE CAN SEE THAT	9
4	WE HAVE TO	8
5	WE CAN ALSO	7
6	WE CAN SAY THAT	6
7	WE CAN SAY	6
8	WE NEED TO	5
9	WE FIND THAT THE	5
10	IN CONCLUSION WE	5

Table 7.4 N-grams containing *we* in Economics

From Table 7.4 and from examining concordance lines, it seems that *we* is used primarily to guide the reader through the writing (e.g. 'Also looking at table 4, *we can see* other variables that have an impact…'), and as representative (e.g. 'To study the relationship between the marginal propensity to consume and the investment multiplier, *we have to* invest the flow of income …').

Engineering

In Engineering, the majority of the keywords are again concerned with ideational content; categories include measurement generally (*measurement, measured, measuring*); experiments concerned with speed measurement (*drag*, *temperature, efficiency, resistance, heat*); and motor mechanics (*cantilever, mechanical, shaft, gauge, motor, cylinder*). Numbers are also key in Engineering (as with Biology) and are again used in calculations, naming visuals, explanations of formulae, and so on (the fact that numbers are key in two disciplines does not indicate an inadequacy in the keyword procedure: there are 12 disciplines in the reference corpus and each discipline was measured against the other 11).

The remaining sections of Chapter 7 discuss the key categories from Chapter 5 in terms of how these indicate differences between the student groups in each discipline (excepting informal items since no link between these and disciplinarity was found). The first key category to be discussed is that of connectors.

7.2.3 Connectors

Chapter 5 established that Chi123 contains a group of preferred connectors (e.g. *on the other hand, meanwhile*). Chapter 6 found that these sequences are used less often in year 3 than years 1/2, suggesting that Chinese students may reduce their use of them over time. This section discusses the connectors used by each student group in Biology, Economics and Engineering.

Within Biology, the use of connectors was found to be in line with that in the corpus overall. Economics and Engineering, however, showed some variation. In Economics, selected keywords from Appendix G suggest that both Chinese and English groups make high use of connectors of logical relations between parts of the text (e.g. *therefore, hence, thus*). Searches for connectors of result/inference (e.g. *therefore, consequently,* thus, hence) and contrast/concession (e.g. *on the other hand, in contrast, alternatively, however,* yet) indicate that each student group prefer particular connectors (Table 7.5).

	Chi-Economics	**(raw)**	**Eng-Economics**	**(raw)**
however	14	(55)	34****	(178)
therefore	17	(65)	23*	(122)
hence	8	(32)	16***	(86)
thus	12****	(44)	4	(20)
in contrast	2**	(7)	0	(1)

Table 7.5 Connectors in Economics per 10,000 words[33] (comparison across student groups, * p<.05; ** p<.01; ***p<.001; ****p<.0001)

There are also differences across the student groups. While texts in Chi-Economics make significantly greater use of *thus* and *in contrast* (though the raw count for the latter is just seven occurrences), texts in Eng-Economics employ *however, therefore* and *hence*. The use of *however* and *therefore* accords with the overall preference in Eng123 for these

33 Note that figures in this chapter are given per 10,000 words, rather than per million words as in the rest of the thesis; this is due to the reduced size of the corpora in the chapter.

connectors; the remaining connectors seem more particular to the student writing in Economics.

Fewer connectors were found in Engineering for both student groups. The only sequence found to display significant difference across the student groups was *on the other hand;* this is used significantly more (*p*=.0001) by the Chinese students than the English students, though the raw counts are low (9 occurrences in Chi-Engineering and zero instances in Eng-Engineering, despite this being the largest discipline subcorpus with a wordcount of 203,370).

7.2.4 First person pronouns

Chapters 5 and 6 discussed differences in the use of first person pronouns between the student groups, and this section examines the use of *I* and *we* in the discipline subcorpora. Initial frequency searches point to differences between the disciplines and also between the student groups (Table 7.6).

per 10,000 words	Chi Biol	Eng Biol	Chi Engin	Eng Engin	Chi Econ	Eng Econ
I	1	1	3	10****	9	14*
we	6	7	17	15	29	23
Total	7	8	20	25	38	37

Table 7.6 Statistical comparison of first person pronouns as used by each student group, i.e. *I* in Chi-Engineering compared to *I* in Eng-Engineering (using log likelihood, * p<.05; ****p<.0001)

Hyland (2001) found that professional academic writers in Science and Engineering disciplines prefer the first person plural over the first person singular; in all three disciplines considered here, both student groups favour *we* over *I*. This could in part be due to the collaborative nature of either student groupwork or professional research in these areas; that is, the plural pronoun is congruent with the multiple researchers or students. However, whereas professional academic writers tend to write joint papers from their collaborative

research, students are more likely to carry out work as a group and then write this up individually (all assignments in the student corpora are single-authored).

Table 7.6 indicates disciplinary differences in the total use of first person pronouns with Biology making the lowest use of these, Engineering taking the median position and Economics the highest. Pronoun use is significant in two discipline subcorpora with Eng-Engineering and Eng-Economics using *I* more frequently than Chi-Engineering and Chi-Economics respectively. Collocates of *I* in Eng-Engineering (the most significant difference between the student groups) include *have, would, was, will, believe, think, had, feel, can, found*, suggesting that *I* is used to report procedures and in reflective sections of writing.

Gardner (2008) and Thompson (2009) comment on the variety of genres contained within Engineering writing. According to Heuboeck et al.'s (2008) classification, Eng-Engineering contains methodology recounts, explanations, and also of 'reflective recounts' (within the narrative recount genre family); the latter group accounts for much of the high use of the first person singular in this subcorpus. For example:

(9) … I don't think this is what a professional engineer is, although I do think that a professional engineer must work in this 'professional manner' … (0354f).

(10) With hindsight I would have called everybody an hour before the meeting to make sure they were coming as I eventually had to later on in the project (0342c).

In summary, as with the use of connectors, the discipline keywords in this study reveal that there are both disciplinary and L1 differences in the use of first person pronouns in Biology, Economics and Engineering. The next section considers the use of visuals and lists across the three disciplines and across student groups.

7.2.5 Visuals and lists

This section concentrates on the use of tables, figures, images and diagrams (collectively referred to as 'visuals') and the use of writing formatted as lists; these features were revealed to be of interest through the keyword analysis in Chapter 5. Counts of tagged

visuals and lists in the corpora in Chapter 5 confirmed this difference, and analysis of the two student corpora by year groups in Chapter 6 suggested that disparity in the use of listlikes in particular become more pronounced over the three years of undergraduate study.

To determine the different usage of visuals and lists across disciplines, these tagged features were counted for each of the six subcorpora (Table 7.7).

	Tables	**Figures**	**Lists**	**Listlikes**
Chi-Biology	15****	25****	1	4
Eng-Biology	5	13	2	6
Chi-Economics	1	14****	2*	25****
Eng-Economics	0	12	1	3
Chi-Engineering	10*	21	7	53****
Eng-Engineering	7	21	10	24

Table 7.7 Use of tables, figures, lists and listlikes per 10,000 words (statistical differences are shown between student groups within each discipline, using log likelihood, * p<.05; ** p<.01; ***p<.001; ****p<.0001).

Table 7.7 suggests that, as with the use of pronouns, there are both disciplinary differences and also differences between the two student groups. Across the three disciplines, Biology and Engineering make more use of tables and figures than Economics, Engineering contains more lists and more listlikes. Within the student groups, for most categories the Chinese groups use each feature more than the English group in the same discipline, and in some cases this is of high statistical significance.

Disciplinary differences in these features are to be expected, since for example Biology entails the understanding of images of natural phenomena (which are tagged as 'figures' in BAWE), and Economics may involve reports with list type writing, yet it is less clear why the student groups should also differ in their usage of these features. In order to gain insight into the way in which these features are used by each student group, 7.3 examines visuals and lists in pairs of texts within each discipline.

7.3 Visuals and lists: Whole text analysis

Thus far, the use of visuals and lists in assignments has been indicated through the extraction of lexical chunks referring to visuals (e.g. *as shown in table*), the use of numerals indicating a list format (5.2, 6.2), and quantified through the counting of the total number of tables, figures, or lists used within a discipline by each student group (7.2). Counting visuals and lists, however, is an approximate measure, since the number of occurrences of a feature does not reveal differences in length, complexity or manner of usage. From the discussion in 7.2, it seems that both discipline and L1 are variables affecting the use of these semiotic resources, and this section aims to begin the process of disentangling these variables, drawing on recent work in multimodal analysis, as discussed in 4.4. Pairs of texts were examined to explore the ways in which visuals and lists are employed through the writing of different students. In each case, the pairs were selected by virtue of answering the same assignment question within the same module at the same university, though the texts appear to be typical of the texts within each L1 and discipline group.

7.3.1 Visuals and extended captions in Biology

The importance of visuals in Biology is emphasized by Dinolfo et al. (2007) in their study of how students 'see' or 'read' cells under a microscope and subsequently describe them. Drawing on Kress (2004), they discuss the '"all-at-once" processing of complex and often competing visual data' and compare this with the 'linear "one-at-a-time" processing that occurs when we read written text line by line' (Dinolfo et al., 2007: 401). For Biology students, then, producing assignments which discuss visual data involves decisions as to which and how many visuals to include, and how to integrate these with the prose of the assignment.

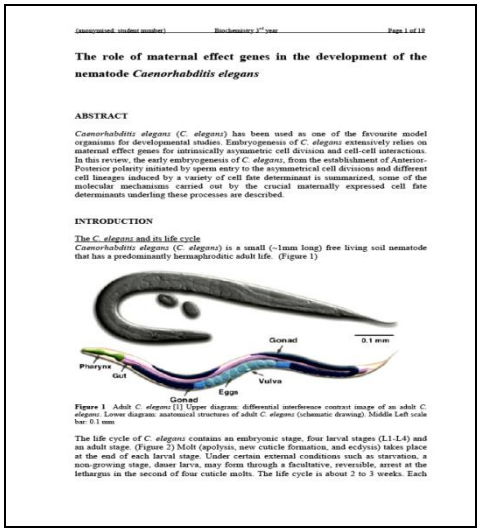
The two texts discussed in this section were written by year 3 Biology students from the same university and from the same module entitled 'Development'; each text was awarded over 70%. The title of each assignment is 'The role of maternal effect genes in the development of the nematode Caenorhabditis elegans', and the assignment requires students to describe and explain the role of these genes on the organism's development;

each assignment is classified in the BAWE corpus data as within the 'explanation' genre. Both texts contain an abstract, and are written in a similarly impersonal style with little overt interpersonal language (though the language of the assignments is not the main focus of this section). The two assignments have approximately the same wordcount when counted through WordSmith Tools after the removal of visuals and their captions. The high use of tables and figures by the Chinese writer increases the number of A4 pages to 15.5 (excluding 3.5 reference or blank end pages), compared to the nine employed by the English writer (though the use of two columns by this writer condenses the prose) (see Table 7.8).

Text feature	**Chinese, text 0434a**	**English, text 0067b**
No. of pages excluding refs	15.5	9
No. of tokens (in WS)	3234	3201
No. of tables	2	0
No. of figures	17	5
Visuals as proportion of whole text	48% (7.5 pp)	22% (2pp)
Layout	whole page	2 columns

Table 7.8 Comparison of two Biology assignments

The category of 'figures' includes images and cross-sectional diagrams of the organism, and process diagrams of its lifecycle and the reproductive process (all of which are labelled as 'figures' by the two students). Figures for both the Biology assignments are mainly in colour, though may have been printed out in black and white for submission to the tutor (both students refer in their captions to colour sections of their visuals, suggesting that colour assignments may have been submitted either online or in print). Both assignments show a sophisticated level of command of the word processing packages used, employing multiple fonts, text wrapping (text 0434a) and dual columns (text 0067b). Figure 7.1 shows the beginning of each assignment, with title, abstract, and introduction.



(anonymised: student number) Biochemistry 3rd year Page 1 of 10

The role of maternal effect genes in the development of the nematode *Caenorhabditis elegans*

ABSTRACT

Caenorhabditis elegans (*C. elegans*) has been used as one of the favourite model organisms for developmental studies. Embryogenesis of *C. elegans* extensively relies on maternal effect genes for intrinsically asymmetric cell division and cell-cell interactions. In this review, the early embryogenesis of *C. elegans*, from the establishment of Anterior-Posterior polarity initiated by sperm entry to the asymmetrical cell divisions and different cell lineages induced by a variety of cell fate determinant is summarized, some of the molecular mechanisms carried out by the crucial maternally expressed cell fate determinants underling these processes are described.

INTRODUCTION

The *C. elegans* and its life cycle

Caenorhabditis elegans (*C. elegans*) is a small (~1mm long) free living soil nematode that has a predominantly hermaphroditic adult life. (Figure 1)

Figure 1 Adult *C. elegans* [1] Upper diagram: differential interference contrast image of an adult *C. elegans*. Lower diagram: anatomical structures of adult *C. elegans* (schematic drawing). Middle Left scale bar: 0.1 mm

The life cycle of *C. elegans* contains an embryonic stage, four larval stages (L1-L4) and an adult stage. (Figure 2) Molt (apolysis, new cuticle formation, and ecdysis) takes place at the end of each larval stage. Under certain external conditions such as starvation, a non-growing stage, dauer larva, may form through a facultative, reversible, arrest at the lethargus in the second of four cuticle molts. The life cycle is about 2 to 3 weeks. Each

Text 0434a, Chinese writer

(anonymised: student name), (anonymised: student number)

The role of maternal effect genes in the development of the nematode *Caenorhabditis elegans*

Maternally expressed genes are essential for the correct patterning and cell-fate determination in the early *Caenorhabditis elegans* embryo. The PAR proteins and MEX-5/6 are responsible for initial polarisation of the zygote, *skn-1* is required to specify the EMS fate, the bifunctional protein PIE-1 is required to maintain the totipotent germ cell lineage and specification of the AB lineage involves a system homologous to Notch in *Drosophila*. This review describes the current understanding of these molecular mechanisms in the specification of cell fates in the pregastrulation embryo.

Introduction

The potential of *Caenorhabditis elegans* as a model organism for the study of embryology emerged in the 1970s (Brenner, 1974). This free-living soil nematode is ideal for studying in the laboratory as it has a rapid period of embryogenesis (16 hours) and each worm has an invariant cell lineage, with exactly 959 somatic cells in the adult, which can be easily traced during development through the transparent cuticle (Sulston & Horvitz, 1977). *C.elegans* is a small round-worm, approximately 1 mm long, that lives for 2-3 weeks and can be fed on *Escherichia coli*, which allows large numbers to be conveniently raised in a Petri dish. The predominant adult form is hermaphroditic, containing both sperm and eggs and therefore reproduction is rapid, either by self-fertilization or by cross-fertilization with the rare males.

The genetics for *C.elegans* is advancing rapidly. It has a small genome at 8×10^7 bp and relatively few genes for a eukaryote – around 17,500. It was the first multicellular organism for which the genome was completely sequenced (*C.elegans* Sequencing Consortium 1999) and approximately 8,000 *C.elegans* proteins have already been matched to homologous human gene transcripts (Lai *et al.*, 2000). Specific mutants may be produced by targeted deletion through transposon insertion or mutagens. Embryos may be manipulated by transformation or injection with transgenes and marker proteins such as green fluorescent protein (GFP) are easily visualised in the transparent embryos. RNA interference (RNAi) is a particularly useful technique for studying maternal effect genes by eliminating the expression of specific maternal or zygotic genes in offspring.

Reproduction

In hermaphrodite worms, fertilization occurs in the spermatheca – an organ where the sperm is stored – when mature oocytes pass from the ovary towards the vulva (Fig 1A-B). The point of sperm entry determines the posterior end of the embryo. After fertilization, a rigid, ovoid-shaped chitin eggshell called the chorion is made (Kemphues & Strome, 1997) and the long axis of this ovoid is termed the anteroposterior (a/p) axis of the embryo.

1

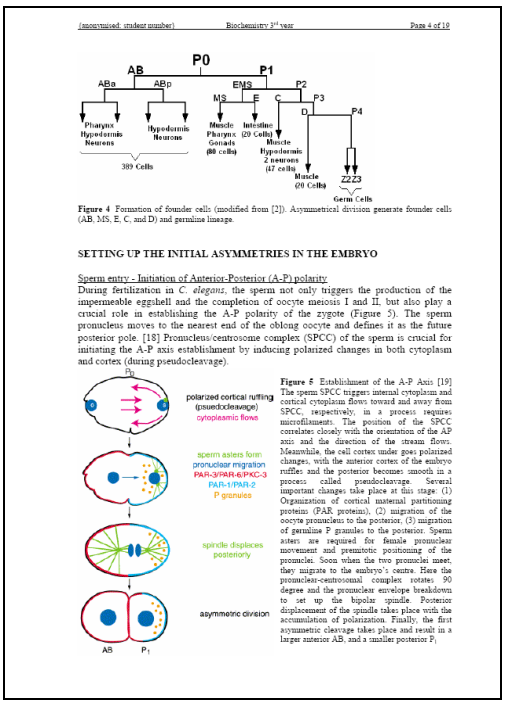
Text 0067b, English writer

Figure 7.1 Page 1 of Biology assignments

The layout constructed by each student is distinctive: the Chinese writer mixes prose with images, tables, and complex, diagrammatic visuals and is perhaps modelled on a textbook. In contrast, the English writer uses two columns and fewer (though still varied) visuals, perhaps emulating the style of a journal article. Texts 0434a and 0067b contain similar quantities of descriptive and explanatory prose (wordcounts of 3234 and 3201 respectively), the difference being that text 0434a more often extends these with illustrative figures. For example in a section headed ‘C. elegans maternal notch system’ there are approximately 350 words of text, including captions, and three figures in text 0434a. In the equivalent section in 0067b, there are approximately 400 words of text and no figures.

Both writers appear to integrate visuals effectively with their writing, using them to both support and extend their prose descriptions. As such, the difference in use is in the extent to which visuals are employed by the Chinese writer, rather than any apparent difference in the nature of this usage. Some of the visuals in the two texts are labelled as adapted from secondary references; others do not contain references though this does not necessarily mean they are devised by the student. Thus, the effort on the student’s part is primarily directed towards writing the surrounding text, rather than devising or reframing the visual. Frequently, particularly for text 0067b, this surrounding text comprises a paragraph of over

100 words of explanatory prose within the caption font; however captions are omitted in BAWE tagging, rendering the comparative wordcounts achieved through WordSmith less accurate (see Figure 7.2 for example of an extended caption from text 0434a).



The extended caption on the bottom right of the page is 186 words long and, after the initial label ('Establishment of the A-P Asix'), is constructed in full sentences and in the same, neutral stance as the rest of the text. As this paragraph of text is formatted as a caption (in a smaller font), it is not included in the text file for this assignment and hence is excluded from WordSmith analysis.

Figure 7.2 Diagrams and extended caption in 0434a p.4 (Chinese writer)

The extended caption given in Figure 7.2 describes the process illustrated graphically by the diagram on the left hand side; however, the words and image are not integrated and the caption paragraph could stand alone. This contrasts with Archer's (2006: 457) exploration of visuals and captions in posters by first year Engineering students in a South African university. Archer viewed students' use of captions as 'free[ing] them up from distanced and neutral academic discourse'. A second example from 0434a shows an extended caption, again written in the same academic style as the rest of the text, and consisting of an integrated description, containing several bracketed references to the visual (e.g. 'shown by arrows', 'yellow lines') (Figure 7.3).

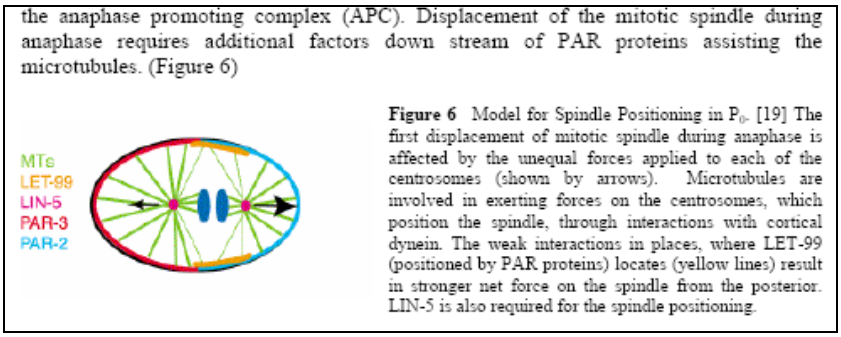


Figure 7.3 Visual and extended caption in 0434a p.5 (Chinese writer)

For both Figures 7.2 and 7.3, the visual, immediate labelling of parts (in colour), and short caption title are copied (and referenced) from the same source article[34]; in both cases the extended caption contains different text to the article caption, suggesting that this is constructed by the student and thus should be treated as part of their prose writing. The source article also uses extended captions which, as with the students' captions, appear to function in parallel with the main prose in that the reader has to break away from the primary text in order to focus on the visuals and then return. The English student similarly employs extended captions to describe and explain the visuals; however, these are more sparsely used than for the Chinese writer; thus, the difference in use of extended captions by the two students appears to be quantitative rather than qualitative.

In a discussion of the effect of using multiple modes, Kress (2009: 54) asks whether different modes are 'merely a kind of duplication of meanings… maybe as "illustration" or "ornamentation"' or whether they are 'distinct, "full" meanings'. In this case study, the visuals used by these Biology students could be simply duplicating the text (whether the extended captions or the main text), they could be extending the meaning provided through language (an option not given by Kress in this quotation), or they could be providing an alternative 'full' way of meaning making. It is difficult, however, to argue for any of these options as a non-specialist in the discipline, and further research as to the effectiveness of visuals (and the use of extended captions) would necessitate the involvement of both lecturer and students in explaining the purpose and integration of the selected resources.

34 The source journal article is: Lyczak, R., Gomes, J.-E., & Bowerman, B. (2002). Heads or Tails: Cell Polarity and Axis Formation in the Early Caenorhabditis elegans Embryo. *Developmental Cell, 3*(2), 157-166.

7.3.2 Bulleted lists vs. connected prose in Economics

The two typical texts in Economics are again from the same university, same module ('Econometrics 1') and have the same title ('Assignment 1'). In line with the mean average wordcounts in Economics given in Table 7.1, the assignment by the Chinese student is around 500 words shorter than the English student's text (Table 7.9). Each assignment was awarded over 70%.

Text feature	Chinese, 0155a	English, 0202j
No. of pages excluding refs[35]	6	6
No. of tokens (in WS)	3731	4242
No. of formulae	19	6
No. of lists	2	0
No. of listlikes	28	0
Lists and listlikes as % of whole text	90%	0%

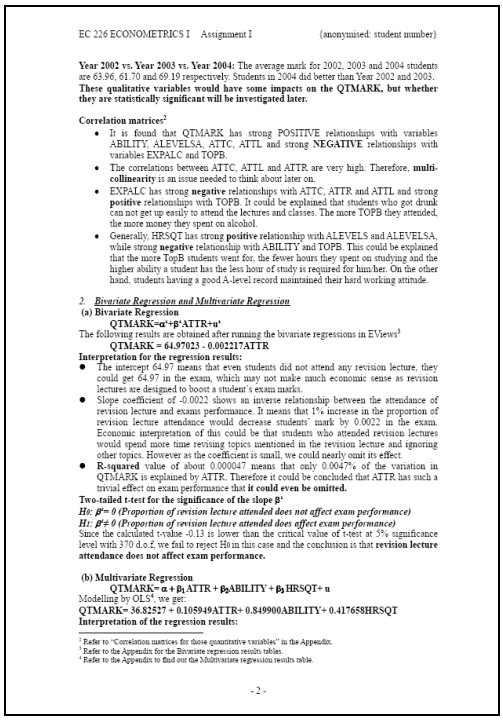
Table 7.9 Comparison of two Economics assignments

The assignment topic concerns the analysis of exam marks from three undergraduate cohorts, and consists of a series of six questions on data tables provided (note that neither student reproduces these data tables in the body of their assignment); the BAWE corpus data lists each assignment as a 'problem question'. The first question of the assignment asks for the main features of the data in a table of descriptive statistics. This description is followed by sections requiring students to carry out tests such as regression analyses on the data (questions 2 - 5). Finally, students are asked to construct a model of exam performance, drawing on their answers to previous sections (question 6).

The most notable overall difference on reading through the two texts is the layout. The Chinese student's work consists almost entirely of a series of lists and listlikes, organized under the six question headings, whereas the English student's assignment is written in connected prose (with some use of formulae given on separate lines), again beneath the six

35 Both assignments contain appendices, though as these were not submitted to BAWE by one of the students, they are not included in the current analysis.

question headings. The example pages in Figure 7.4 illustrate the use of bulleted lists and continuous prose.



EC 226 ECONOMETRICS I Assignment I {anonymised: student number}

Year 2002 vs. Year 2003 vs. Year 2004: The average mark for 2002, 2003 and 2004 students are 63.96, 61.70 and 69.19 respectively. Students in 2004 did better than Year 2002 and 2003. **These qualitative variables would have some impacts on the QTMARK, but whether they are statistically significant will be investigated later.**

Correlation matrices[2]

- It is found that QTMARK has strong POSITIVE relationships with variables ABILITY, ALEVELSA, ATTC, ATTL and strong **NEGATIVE** relationships with variables EXPALC and TOPB.
- The correlations between ATTC, ATTL and ATTR are very high. Therefore, **multi-collinearity** is an issue needed to think about later on.
- EXPALC has strong **negative** relationships with ATTC, ATTR and ATTL and strong **positive** relationships with TOPB. It could be explained that students who got drunk can not get up easily to attend the lectures and classes. The more TOPB they attended, the more money they spent on alcohol.
- Generally, HRSQT has strong **positive** relationship with ALEVELS and ALEVELSA, while strong **negative** relationship with ABILITY and TOPB. This could be explained that the more TopB students went for, the fewer hours they spent on studying and the higher ability a student has the less hour of study is required for him/her. On the other hand, students having a good A-level record maintained their hard working attitude.

2. ***Bivariate Regression and Multivariate Regression***

(a) Bivariate Regression

QTMARK=α'+β'ATTR+u'

The following results are obtained after running the bivariate regressions in EViews[3]

QTMARK = 64.97023 - 0.002217ATTR

Interpretation for the regression results:

- The intercept 64.97 means that even students did not attend any revision lecture, they could get 64.97 in the exam, which may not make much economic sense as revision lectures are designed to boost a student's exam marks.
- Slope coefficient of -0.0022 shows an inverse relationship between the attendance of revision lecture and exams performance. It means that 1% increase in the proportion of revision lecture attendance would decrease students' mark by 0.0022 in the exam. Economic interpretation of this could be that students who attended revision lectures would spend more time revising topics mentioned in the revision lecture and ignoring other topics. However as the coefficient is small, we could nearly omit its effect.
- **R-squared** value of about 0.000047 means that only 0.0047% of the variation in QTMARK is explained by ATTR. Therefore it could be concluded that ATTR has such a trivial effect on exam performance that **it could even be omitted.**

Two-tailed t-test for the significance of the slope β'

H0: β'= 0 (Proportion of revision lecture attended does not affect exam performance)

H1: β'≠ 0 (Proportion of revision lecture attended does affect exam performance)

Since the calculated t-value -0.13 is lower than the critical value of t-test at 5% significance level with 370 d.o.f, we fail to reject H0 in this case and the conclusion is that **revision lecture attendance does not affect exam performance.**

(b) Multivariate Regression

QTMARK= α + β_1ATTR + β_2ABILITY + β_3HRSQT+ u

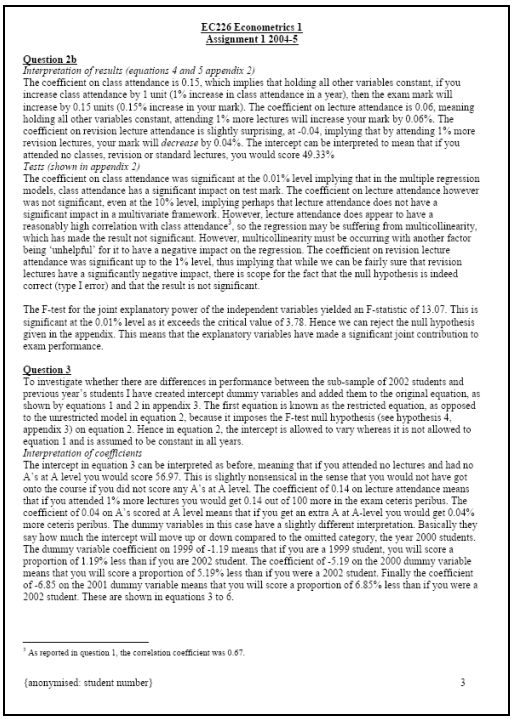
Modelling by OLS[4], we get:

QTMARK= 36.83527 + 0.105949ATTR+ 0.849900ABILITY+ 0.417658HRSQT

Interpretation of the regression results:

[2] Refer to "Correlation matrices for those quantitative variables" in the Appendix.
[3] Refer to the Appendix for the Bivariate regression results tables.
[4] Refer to the Appendix to find out the Multivariate regression results table.

- 2 -



EC226 Econometrics 1
Assignment 1 2004-5

Question 2b

Interpretation of results (equations 4 and 5 appendix 2)

The coefficient on class attendance is 0.15, which implies that holding all other variables constant, if you increase class attendance by 1 unit (1% increase in class attendance in a year), then the exam mark will increase by 0.15 units (0.15% increase in your mark). The coefficient on lecture attendance is 0.06, meaning holding all other variables constant, attending 1% more lectures will increase your mark by 0.06%. The coefficient on revision lecture attendance is slightly surprising, at -0.04, implying that by attending 1% more revision lectures, your mark will *decrease* by 0.04%. The intercept can be interpreted to mean that if you attended no classes, revision or standard lectures, you would score 49.33%

Tests (shown in appendix 2)

The coefficient on class attendance was significant at the 0.01% level implying that in the multiple regression models, class attendance has a significant impact on test mark. The coefficient on lecture attendance however was not significant, even at the 10% level, implying perhaps that lecture attendance does not have a significant impact in a multivariate framework. However, lecture attendance does appear to have a reasonably high correlation with class attendance[3], so the regression may be suffering from multicollinearity, which has made the result not significant. However, multicollinearity must be occurring with another factor being 'unhelpful' for it to have a negative impact on the regression. The coefficient on revision lecture attendance was significant up to the 1% level, thus implying that while we can be fairly sure that revision lectures have a significantly negative impact, there is scope for the fact that the null hypothesis is indeed correct (type I error) and that the result is not significant.

The F-test for the joint explanatory power of the independent variables yielded an F-statistic of 13.07. This is significant at the 0.01% level as it exceeds the critical value of 3.78. Hence we can reject the null hypothesis given in the appendix. This means that the explanatory variables have made a significant joint contribution to exam performance.

Question 3

To investigate whether there are differences in performance between the sub-sample of 2002 students and previous year's students I have created intercept dummy variables and added them to the original equation, as shown by equations 1 and 2 in appendix 3. The first equation is known as the restricted equation, as opposed to the unrestricted model in equation 2, because it imposes the F-test null hypothesis (see hypothesis 4, appendix 3) on equation 2. Hence in equation 2, the intercept is allowed to vary whereas it is not allowed to equation 1 and is assumed to be constant in all years.

Interpretation of coefficients

The intercept in equation 3 can be interpreted as before, meaning that if you attended no lectures and had no A's at A level you would score 56.97. This is slightly nonsensical in the sense that you would not have got onto the course if you did not score any A's at A level. The coefficient of 0.14 on lecture attendance means that if you attended 1% more lectures you would get 0.14 out of 100 more in the exam ceteris peribus. The coefficient of 0.04 on A's scored at A level means that if you get an extra A at A-level you would get 0.04% more ceteris peribus. The dummy variables in this case have a slightly different interpretation. Basically they say how much the intercept will move up or down compared to the omitted category, the year 2000 students. The dummy variable coefficient on 1999 of -1.19 means that if you are a 1999 student, you will score a proportion of 1.19% less than if you are 2002 student. The coefficient of -5.19 on the 2000 dummy variable means that you will score a proportion of 5.19% less than if you were a 2002 student. Finally the coefficient of -6.85 on the 2001 dummy variable means that you will score a proportion of 6.85% less than if you were a 2002 student. These are shown in equations 3 to 6.

[3] As reported in question 1, the correlation coefficient was 0.67.

{anonymised: student number} 3

Text 0155a, Chinese writer Text 0202j, English writer

Figure 7.4 Example page of each Economics assignment

Lea and Street (2006: 372) highlight the effect of different layouts, suggesting that 'writing always creates meaning through layout, as well as through the use of words'. For instance the bulleted lists of text 0155a give a sense of information items being provided cumulatively, in contrast to the connected prose of text 0202j (Figure 7.4). The bullet points and wide margins of the former also provide more white space, highlighting each point made, unlike the narrow margins of the latter.

The Chinese writer's text (0155a) is set out as a report, with a brief, 3-line 'introduction' followed by headings consisting of the question number and a meaningful title (e.g. '1. Main features of the data'). Within each of these sections, points are given as a bulleted list (see Figure 7.5); frequently this list includes further items given within lists. On occasion, list items are not separated by bullets or similar formatting (and so would not be included in the BAWE listlikes count), but consist simply of a series of sentences which each begin on a new line.

The text is broken up throughout with the writer making use of italics or emboldening to highlight lines containing formulae; bold is also used for subheadings and key concluding points (Figure 7.5).

Comparisons of data across various groups

Pure Economics degree vs. non pure Economics degree: Students doing pure Economics degree scored 66.23 on average, while students doing a mixed-Economics degree scored 61.68 (very significant).

Female vs. Male: The average female students got 63.8, compared to 65.4 for male students.

UK students vs. non-UK students: On average, UK students gained 64.63 while non-UK students gained 66.38.

Number of parents who attended university: Those students whose parents never attended university achieved 64.12 on average, those with one parent attended university achieved 64.99 averagely, and those with both parents attended university achieved 65.59 on average.

Figure 7.5 Extract from text 0155a (Chinese writer)

The high use of lists throughout this assignment serves to organize the text, reducing the need for connecting words and phrases. For example in Figure 7.5, the possible ways of grouping the data are given as headings (e.g. 'male vs. female'); this use of meaningful headings and the fragmented layout removes the need for linking phrases such as 'The next group considered is'. Sometimes, the reader has to interpret the relevance of each piece of information in a list; for example in the final list item of Figure 7.5 we have to compare the numbers given to extract the intended meaning that students with two parents who attended university achieved a higher score than those with one or none (cf. Hinds', 1987, notion of reader/writer responsibility). The final paragraph to this assignment is headed 'comments on the model' and impartially restates the main points and the evidence on which these are based. A more personal note is conveyed in the penultimate section to question 6 in which the writer says 'I am concluding to the best of my ability, that the optimal function should be of a linear form'.

In contrast to the lists used in 0155a, text 0202j consists largely of paragraphs of sentences linked with *however, unfortunately, in all cases* and so on. There is a discursive introduction which sets out the rest of the text: 'in this project I shall firstly analyze various factors… I

shall then regress and analyze… before concluding….'. The main headings relate to the assignment outline (e.g. 'Question 1'); within these, limited use is made of subheadings. The prose is also punctuated by occasional formulae. However, the overall effect is of a continuous piece of writing with signposting used judiciously to guide the reader through (see extract in Figure 7.6).

> To make some comments about these results, we need to break this up into sub-samples. Firstly we can break it up according to sex, as Siegfied and strand did. For males, table 2 shows that the mean score is 65.4%, which is higher than the corresponding score for females of 63.75%. This agrees with Siegried and Strand's paper which claims males do better than females. However the standard deviation for males is lower than for females, 12.64% compared to 14.02%.

Figure 7.6 Extract from text 0202j

A comparison of Figure 7.5 and 7.6 illustrates the more discursive nature of 0202j. The English writer of text 0202j first introduces the need for 'sub-samples' within the results, and then guides the reader through these ('firstly we can break it up according to…'). Figure 7.6 also shows how the writer of 0202j broadens the discussion out from the given data, making links with other texts ('This agrees with Siegfried and Strand's paper…'). While text 0202j brings in additional voices through citation, it displays a more minimal approach than 0155a to providing background information for calculations (6 formulae are given by the English writer compared to 19 provided by the Chinese writer).

The signposting shown in Figure 7.6 continues in later sections of 0202j; for example 'given my results from previous questions and previous research, I am now going to try and formulate a model of exam performance…' and 'let us turn back to graph 3 of appendix 1'. The final paragraph of this assignment begins 'in conclusion' and reiterates the writer's views on the model. Unlike the writer of text 0155a, the English writer of 0202j restates the main results while also conveying the achievement of the assignment: 'I have calculated a model of exam performance based on a number of variables, as given above'. The writer relates these variables to 'theory from economists such as Romer' and also to their 'own experience', whereas the Chinese writer restricts their description to the provided data. Overall, the Chinese writer gives their responses to the six questions within the assignment

succinctly and with little metatext or comment but focuses instead on the calculations. In comparison, the English writer guides the reader through the work and reflects more on the process. This disparity has similarities with Mauranen's (1994) work on Finnish and Anglo-American writers' strategies in addressing the reader, in which she examines the difficulties of Finnish exchange students meeting the challenge of studying in the new environment of a UK university. Mauranen found that the Finnish students' texts contained *less* explicit reader guidance than texts written by UK students. While cultural differences exist between the UK and Finland, the gap between study discourses or genres of the UK and China is likely to be even greater.

7.3.3 Formulae and white space in Engineering

The pairs of texts in Engineering are from the same university and module, both from year 2 students, and are each entitled 'centrifugal pump experiment', and again appear typical for each student group (Table 7.10).

Text feature	Chinese, 0254g	English, 0329e
No. of pages excluding refs	11	5.5
No. of tokens (in WS)	1,432	2,064
No. of tables	1	0
No. of images	1	0
No. of formulae[36]	34	10
No. of lists	2	2
No. of listlikes	9	0
No. of block quotes	0	3

Table 7.10 Comparison of two Engineering assignments

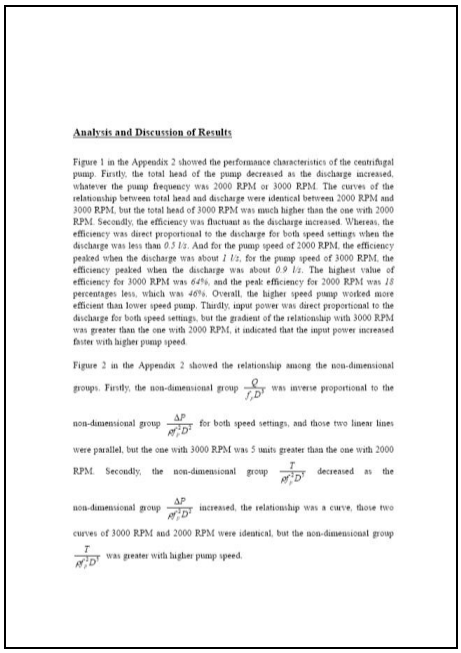
The assignments are labelled by the student as a 'laboratory report' (Chinese) and a 'laboratory exercise' (English), though both are classed in the BAWE data as 'methodology recounts' (the genre title 'laboratory report/exercise' is not used in the BAWE genre

36 The number of formulae for the English text has been altered from the three given in BAWE data to ten, to correct a disparity in the way formulae were tagged in the two texts. The 'figure' in the Chinese text is listed here as an 'image'.

taxonomy). The assignment from the Chinese student achieved a ‘merit’ (i.e. scoring over 60%), the exact grade for the English student is not known.

Both assignments are divided into sections, listed in an initial table of contents and given as headings in the body of each text: for example ‘summary’, ‘introduction’, ‘apparatus and methods’, ‘evaluation of results’. The Chinese writer begins each section on a new page (resulting in white space at the end of several of the pages), whereas the English student simply uses a line break and begins a new section. The most striking aspect of a comparison of the two PDFs is this use of white space. Although the Chinese writer’s assignment (excluding references and appendices) contains twice as many pages in PDF format as the English assignment, the former contains only two-thirds of the wordcount of the latter.

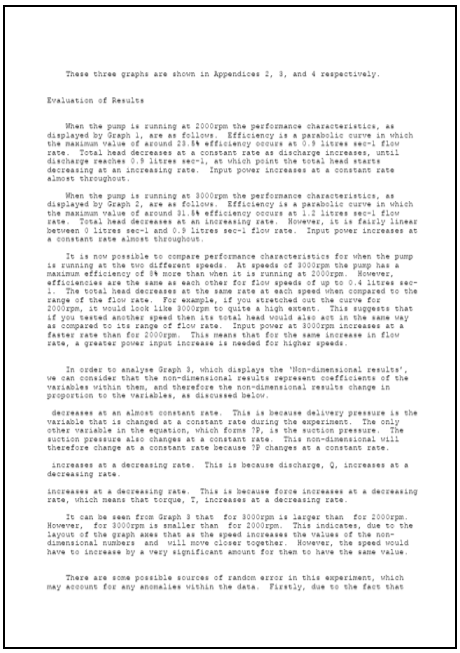
The different use of formulae and prose is illustrated by the pages in Figure 7.7. The Chinese student’s discussion weaves formulae and prose discussion together, whereas the English student’s evaluation is given as a series of short paragraphs.



Analysis and Discussion of Results

Figure 1 in the Appendix 2 showed the performance characteristics of the centrifugal pump. Firstly, the total head of the pump decreased as the discharge increased, whatever the pump frequency was 2000 RPM or 3000 RPM. The curves of the relationship between total head and discharge were identical between 2000 RPM and 3000 RPM, but the total head of 3000 RPM was much higher than the one with 2000 RPM. Secondly, the efficiency was fluctuant as the discharge increased. Whereas, the efficiency was direct proportional to the discharge for both speed settings when the discharge was less than *0.5 l/s*. And for the pump speed of 2000 RPM, the efficiency peaked when the discharge was about *1 l/s*, for the pump speed of 3000 RPM, the efficiency peaked when the discharge was about *0.9 l/s*. The highest value of efficiency for 3000 RPM was *64%*, and the peak efficiency for 2000 RPM was *18* percentages less, which was *46%*. Overall, the higher speed pump worked more efficient than lower speed pump. Thirdly, input power was direct proportional to the discharge for both speed settings, but the gradient of the relationship with 3000 RPM was greater than the one with 2000 RPM, it indicated that the input power increased faster with higher pump speed.

Figure 2 in the Appendix 2 showed the relationship among the non-dimensional groups. Firstly, the non-dimensional group $\frac{Q}{f_r D^3}$ was inverse proportional to the non-dimensional group $\frac{\Delta P}{\rho f_r^2 D^2}$ for both speed settings, and those two linear lines were parallel, but the one with 3000 RPM was 5 units greater than the one with 2000 RPM. Secondly, the non-dimensional group $\frac{T}{\rho f_r^2 D^5}$ decreased as the non-dimensional group $\frac{\Delta P}{\rho f_r^2 D^2}$ increased, the relationship was a curve, those two curves of 3000 RPM and 2000 RPM were identical, but the non-dimensional group $\frac{T}{\rho f_r^2 D^5}$ was greater with higher pump speed.



These three graphs are shown in Appendices 2, 3, and 4 respectively.

Evaluation of Results

When the pump is running at 2000rpm the performance characteristics, as displayed by Graph 1, are as follows. Efficiency is a parabolic curve in which the maximum value of around 23.5% efficiency occurs at 0.9 litres sec-1 flow rate. Total head decreases at a constant rate as discharge increases, until discharge reaches 0.9 litres sec-1, at which point the total head starts decreasing at an increasing rate. Input power increases at a constant rate almost throughout.

When the pump is running at 3000rpm the performance characteristics, as displayed by Graph 2, are as follows. Efficiency is a parabolic curve in which the maximum value of around 31.5% efficiency occurs at 1.2 litres sec-1 flow rate. Total head decreases at an increasing rate. However, it is fairly linear between 0 litres sec-1 and 0.8 litres sec-1 flow rate. Input power increases at a constant rate almost throughout.

It is now possible to compare performance characteristics for when the pump is running at the two different speeds. At speeds of 3000rpm the pump has a maximum efficiency of 8% more than when it is running at 2000rpm. However, efficiencies are the same as each other for flow speeds of up to 0.4 litres sec-1. The total head decreases at the same rate at each speed when compared to the range of the flow rate. For example, if you stretched out the curve for 2000rpm, it would look like 3000rpm to quite a high extent. This suggests that if you tested another speed then its total head would also act in the same way as compared to its range of flow rate. Input power at 3000rpm increases at a faster rate than for 2000rpm. This means that for the same increase in flow rate, a greater power input increase is needed for higher speeds.

In order to analyse Graph 3, which displays the ‘Non-dimensional results’, we can consider that the non-dimensional results represent coefficients of the variables within them, and therefore the non-dimensional results change in proportion to the variables, as discussed below.

decreases at an almost constant rate. This is because delivery pressure is the variable that is changed at a constant rate during the experiment. The only other variable in the equation, which forms ?P, is the suction pressure. The suction pressure also changes at a constant rate. This non-dimensional will therefore change at a constant rate because ?P changes at a constant rate.

increases at a decreasing rate. This is because discharge, Q, increases at a decreasing rate.

increases at a decreasing rate. This is because force increases at a decreasing rate, which means that torque, T, increases at a decreasing rate.

It can be seen from Graph 3 that for 3000rpm is larger than for 2000rpm. However, for 3000rpm is smaller than for 2000rpm. This indicates, due to the layout of the graph axes that as the speed increases the values of the non-dimensional numbers and will move closer together. However, the speed would have to increase by a very significant amount for them to have the same value.

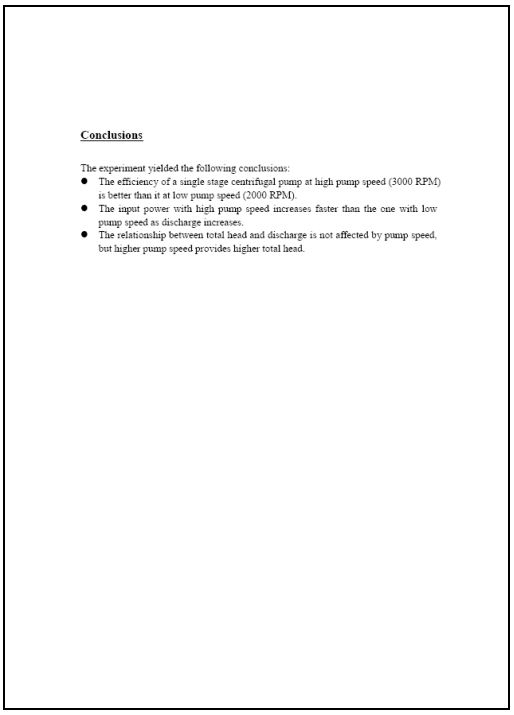
There are some possible sources of random error in this experiment, which may account for any anomalies within the data. Firstly, due to the fact that

Text 0254g p.8 (Chinese) Text 0329e p.5 (English)

Figure 7.7 Discussion/evaluation sections in two Engineering assignments

Throughout the assignment, the Chinese student uses emboldened, underlined headings (as illustrated in Figure 7.7) and employs numbered or bulleted lists and tables to present data. In contrast, the English student does not use any formatting (other than paragraph indentation), giving headings in the same font size with no embellishment. It is unclear as to whether the English student did not know *how* to format their assignment, or whether they simply did not think this was important (e.g. the table of contents is given as a simple list of the headings, without numbering, bulleting or page numbers).

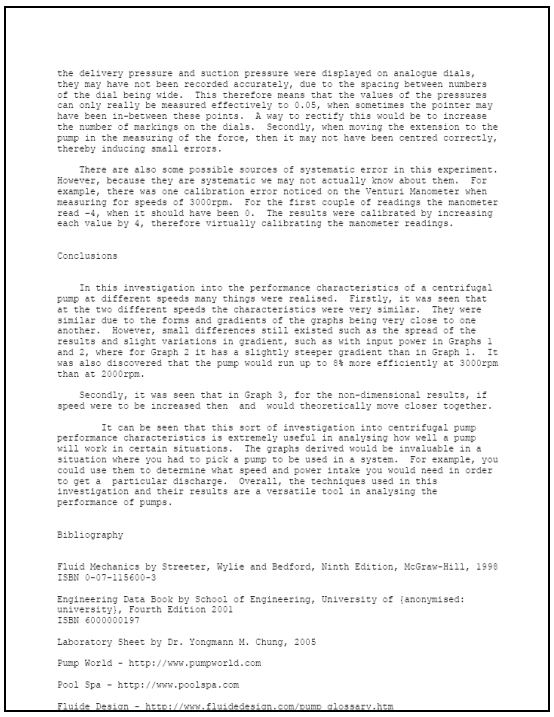
Figure 7.8 gives each student's final section in the report, in each case headed 'Conclusions'; this again illustrates the greater use of white space by the Chinese student.



Conclusions

The experiment yielded the following conclusions:

- The efficiency of a single stage centrifugal pump at high pump speed (3000 RPM) is better than it at low pump speed (2000 RPM).
- The input power with high pump speed increases faster than the one with low pump speed as discharge increases.
- The relationship between total head and discharge is not affected by pump speed, but higher pump speed provides higher total head.



the delivery pressure and suction pressure were displayed on analogue dials, they may have not been recorded accurately, due to the spacing between numbers of the dial being wide. This therefore means that the values of the pressures can only really be measured effectively to 0.05, when sometimes the pointer may have been in-between these points. A way to rectify this would be to increase the number of markings on the dials. Secondly, when moving the extension to the pump in the measuring of the force, then it may not have been centred correctly, thereby inducing small errors.

There are also some possible sources of systematic error in this experiment. However, because they are systematic we may not actually know about them. For example, there was one calibration error noticed on the Venturi Manometer when measuring for speeds of 3000rpm. For the first couple of readings the manometer read -4, when it should have been 0. The results were calibrated by increasing each value by 4, therefore virtually calibrating the manometer readings.

Conclusions

In this investigation into the performance characteristics of a centrifugal pump at different speeds many things were realised. Firstly, it was seen that at the two different speeds the characteristics were very similar. They were similar due to the forms and gradients of the graphs being very close to one another. However, small differences still existed such as the spread of the results and slight variations in gradient, such as with input power in Graphs 1 and 2, where for Graph 2 it has a slightly steeper gradient than in Graph 1. It was also discovered that the pump would run up to 8% more efficiently at 3000rpm than at 2000rpm.

Secondly, it was seen that in Graph 3, for the non-dimensional results, if speed were to be increased then and would theoretically move closer together.

It can be seen that this sort of investigation into centrifugal pump performance characteristics is extremely useful in analysing how well a pump will work in certain situations. The graphs derived would be invaluable in a situation where you had to pick a pump to be used in a system. For example, you could use them to determine what speed and power intake you would need in order to get a particular discharge. Overall, the techniques used in this investigation and their results are a versatile tool in analysing the performance of pumps.

Bibliography

Fluid Mechanics by Streeter, Wylie and Bedford, Ninth Edition, McGraw-Hill, 1998 ISBN 0-07-115600-3

Engineering Data Book by School of Engineering, University of {anonymised: university}, Fourth Edition 2001
ISBN 6000000197

Laboratory Sheet by Dr. Yongmann M. Chung, 2005

Pump World - http://www.pumpworld.com

Pool Spa - http://www.poolspa.com

Fluide Design - http://www.fluidedesign.com/pump_glossary.htm

0254g Conclusions p.12 0329e Conclusions (central section) p.6

Figure 7.8 'Conclusions' in two Engineering assignments

The left-hand box of Figure 7.8 shows how the Chinese student provides a list of three bullet points, given in complete sentences, which state the bald facts of the experiment:

The experiment yielded the following conclusions:

- The efficiency of a single stage centrifugal pump at high pump speed (3000 RPM) is better than it at low pump speed (2000 RPM)
- The input power with high pump speed increases faster than the one with low pump speed as discharge increases.
- The relationship between total head and discharge is not affected by pump speed, but higher pump speed provides higher total head.

(Conclusion section of text 0254g).

In contrast, the English student's conclusion (the central section of the right-hand box) is far more discursive, introducing the results and relating them to the experiments:

> In this investigation many into the performance characteristics of a centrifugal pump at different speeds many things were realised. Firstly, it was seen that at the two different speeds the characteristics were very similar. They were similar due to the forms and gradients of the graphs being very close to one another. However, small differences still existed such as the spread of the results…

(Beginning of Conclusion for text 0329j).

The two Engineering texts illustrate the different use of formulae by the students and the different attention paid to presentation. In particular, the use of formulae interwoven with prose, lists given as a series of bulleted points and new pages for each section in the Chinese student's text results in a generous use of white space.

Summary

The three pairs of assignments considered in this section illustrate the diversity permitted. As all six assignments are deemed proficient by their university tutors (as defined in 2.2), it seems that features such as making extensive use of visuals or formulae, writing in lists or in connected prose are all acceptable. It is difficult to speculate, however, as to the preferred characteristics of student writing in particular disciplines, and the next section draws on Biology, Economics and Engineering lecturers' views on assignment-writing in order to understand what is valued within the disciplines.

7.3.4 Interviews with lecturers

Alongside the collection of assignments for the BAWE corpus, interviews with lecturers were carried out to provide an emic perspective on the texts (as discussed in Nesi and Gardner, 2006; Leedham, 2009). The BAWE lecturer interviews included questions relating to the role

of assignment-writing in the department, valued and 'disliked' features of undergraduate assignments, and the writing needs of NNS students. Of concern in this chapter, are the interviews with lecturers in Biology (n=4), Economics (n=3), and Engineering (n=3) as these shed some light on the different strategies employed by students in these disciplines, particularly with regard to the use of visuals. The following discussion of the interview data is considered under four themes, beginning with overall remarks on assignment structure (as this pertains to layout employed in texts considered in this section), lecturers' stated preference for concise writing, followed by comments relating to the use of visuals and lists. Interviewee comments are given anonymously, and ascribed to the discipline.

Structure

In Biology, the main type of assignment is the report based on a laboratory exercise and written in clear sections with headings such as introduction, materials and methods, results (including graphs), discussion. These reports give students practice in analyzing and interpreting data, and are regarded as 'providing a better indication of student's level of understanding' than an essay since in a report 'there is no existing document out there which explains how to interpret their data' (Biology lecturer). The structure provided by a report format may also be favourably viewed by students: in Engineering, one tutor commented that for students who have 'become accustomed to writing structured reports, the prospect of an essay on professional ethics is daunting'. Despite this, however, students in Engineering are required to write in a multitude of different genres over the course of their undergraduate study, writing essays aimed at non-specialized readers, reflective writing on the experience of groupwork, as well as the ubiquitous laboratory reports (cf. remarks on genres in Engineering by Gardner, 2008, and Thompson, 2009, as discussed in 2.2). In contrast, Economics assignments appear to be mainly essays, which 'over time, … should approximate ever more closely to the writing that academics submit for publication in learned/scientific journals'. Biology students are similarly told that they should 'write in the style of current opinion journals' (cf. the two column-layout of text 0067b in this discipline).

Being concise

Despite the differences in disciplines and assignment genres, lecturers in all three disciplines placed value on the ability to write in a concise way. In Engineering, lecturers value the ability to be 'clear and concise', 'succinct', and point to a dislike of 'verbosity', in Economics one lecturer commented on their favouring of 'precision, incision, concision', and in Biology, a lecturer commented that 'there's never been a penalty for an essay that's too short'. In all three disciplines, it seems that the ability to be concise is favoured over lengthy prose.

Employing visuals

Again, despite disciplinary differences, lecturers in all three disciplines commented on the importance of visuals in assignments. In Biology, one interviewee suggested that a lab report of five or six pages should include 2-4 pages of diagrams, pointing to the visual nature of the discipline (and echoing the argument in Dinolfo et al., 2007, discussed earlier in 7.3.1). In Economics, a 'typical' essay contains both diagrams and formulae 'as the spine of the essay'. In Engineering, marks for presentation may include the assessment of diagrams, tables and overall layout. A Biology lecturer remarked that including visuals helps undergraduates to gain better marks since they can refer to these rather than having to describe everything and risking the introduction of errors. Students are, however, 'incredibly reluctant' to construct their own diagrams and tend to use existing ones taken from websites (Biology). While this is acceptable, it is preferable if students at least devise their own accompanying caption (cf. the Chinese student's reproduced diagrams with original captions). In both Biology and Engineering, mislabelled axes on graphs were mentioned as a particular dislike. Two out of the three interviewees in Economics commented on the challenge involved in analyzing visuals in prose. While the pair of Economics assignments discussed previously do not include visuals, the main task set for them is the interpretation of a dataset given in graphs and other visual forms. For essays including diagrams, the 'challenge' is 'to marry the diagrams with the text', since 'the key writing skill for an economist is the ability to demonstrate in writing about a diagram an understanding of the analysis' it presents (Economics).

Writing in lists

Little mention was made of list writing in the interviews. One Economics lecturer stated a dislike of written work containing 'just diagrams and incomplete notes' rather than complete sentences. However, the lists and listlikes presented in text 0155a (Economics) in this section contain full sentences given as lists, rather than as sentence fragments so would presumably not be viewed in this negative way. An Engineering lecturer similarly remarked that he disliked use of bullet points as a space-saving feature; the use of bulleted lists is perhaps viewed as a way of circumventing the occasional setting of page (as well as word) limits in Engineering assignments.

Finally, one Economics interviewee highlighted the number of L2 English students and staff, commenting that 'you might well find a graduate teaching assistant from Mexico teaching a student from China or Africa' (Economics). This is an important point since assignment markers may have a different L1 and cultural background to both the student and the lecturer.

Summary

The lecturer interviews point to commonalities across the three disciplines, favouring clear, concise writing, and the use of appropriate visuals with sufficient prose to explain them. While writing in a list format is not specifically given as a valued feature, it could be the case that lecturers value this form of brevity when marking. Note that items within 'listlikes' in BAWE contain full sentences (though items in 'lists' may consist of sentence fragments).

7.4 Conclusion and summary

In Chapters 5 and 6, quantitative evidence on the use of the features of informal language, connectors, first person pronouns, visuals and lists was presented, and discussed in relation to the literature on student writing reported in Chapter 2. Chapter 7 has discussed the significance of these characteristics in the disciplines of Biology, Economics and Engineering, suggesting that while no informal and few linking items were found to distinguish the disciplines, the use of first person pronouns, and visuals and lists differ

across both disciplines and L1 groups. In terms of first person pronoun use, it was found that for both student groups Economics texts use first person pronouns the most, Engineering texts take central position, and Biology texts use them the least. The use of *I* in Eng-Engineering was found to be of high statistical significance compared to its use in Chi-Engineering, and it is suggested that this is due to the high use of reflective writing in the English students' writing compared to the writing produced by Chinese students. The use of visuals and lists varies both across disciplines and across student groups, with Chinese students employing more tables, figures and lists than English students. Quantitative findings for the use of these features were followed up with whole text analysis of pairs of assignments from each discipline. It is suggested that employing visuals and lists may be strategies for Chinese students who have to meet the challenges of producing multiple, extended pieces of writing within unfamiliar genres in their L2. As such, these strategies allow students to present their findings and views clearly and effectively, while reducing the quantity of connected prose they have to produce. Presenting through visuals and within lists may also be used as strategies by other L2 English writers, though further research on student writing across L2s and disciplines would be needed to explore this. It may also be the case that the Chinese and English students are increasing the affordances of the available semiotic resources, or perhaps employing different semiotic resources, in their assignment-writing. The different use of resources may even play a part in transforming student academic and disciplinary genres such that the range of acceptable means of responding to an assignment question is extended.

The next chapter draws together the analysis and findings from Chapters 5, 6, and 7, and relates them to earlier findings reported in the literature.

CHAPTER 8 CONCLUSIONS

I began this thesis with the assertion that written assessment is the principal way in which both NS and NNS undergraduate students are judged during their university studies (e.g. Douglas, 2010; Lillis and Scott, 2008; North, 2005b). The literature review carried out in Chapter 2 revealed that despite the importance of assessed undergraduate writing, relatively little research has been carried out. This lack of research is particularly apparent for NNS undergraduate writing, with most studies of L2 writing drawing on corpora of very short essays rather than authentic texts produced in the context of regular assessment tasks. Since Chinese students now comprise the largest NNS cohort of undergraduates in the UK, research into their writing is particularly needed. The primary aim of this thesis was thus to add to the body of knowledge on current undergraduate student writing, through examination of a dataset of Chinese students' assignments submitted to UK universities between 2000 and 2008. The contribution of the study has been in examining features of Chinese and English students' writing and discussing how these assist in meeting the challenges of undergraduate written assessment. Features initially extracted through keyword analysis have been compared across the whole of the Chinese and English corpora used in the study, from year 1/2 to year 3, and also across disciplines, to explore variations in the writing. The thesis has been primarily concerned with describing a variety of academic writing through the extraction of n-grams, rather than exploring the psycholinguistic reality behind retrieved frequency-based chunks. Chunks extracted using keyword searches reveal the differences between corpora, and human intuition is used in grouping the resulting keywords into key categories of items.

In addition to answering the three research questions posed in Chapter 2, I have also questioned the extent to which findings from learner corpora can be extended beyond the short essays which form those corpora. Additionally, I have claimed that corpus linguistic methodology can be most fruitfully used when combined with whole text reading in order to keep the larger context in view and to allow for the multimodal analysis of the texts.

This final chapter brings together findings from throughout the thesis. It is divided into three main sections; in 8.1, I draw together the principal findings from the study, focusing on the three research questions. In 8.2, I discuss the implications of these findings, and in 8.3, I examine limitations of the study and provide suggestions for how these could be overcome in future research.

8.1 Answering the research questions

Findings from the research literature on both Chinese students' writing and more widely on NNS' writing are the high use of particular lexical items and chunks, including informal or 'speech-like' items and connectors, and the use of first and second person pronouns. Most studies assume that any differences between NNS and NS writing are *deficits* on the part of the NNSs; in this study I suggest that NNS and NS student writing are instead different *varieties*. This perspective does not mean that all features of student writing are equally valued by the academy and by disciplinary communities; however, the texts used in this study are successful in that they have been awarded high grades by discipline lecturers. Differences in the writing by the student groups may thus indicate variation in ways of satisfactorily accomplishing the task set by disciplinary specialists. In the thesis, I frame these variations within academic literacies research which emphasizes the difficulties of the writing tasks for both NS and NNS students.

A particular focus of the study has been the exploration of lexical chunks, and how these offer a means of investigating shared patterns in academic writing. In Chapter 3, I discussed the extension of Hoey's (2005) lexical priming theory to a consideration of chunks within student writing, arguing that a discourse community's agreed language patterns result from the shared primings of individuals within the community and the harmonising of these primings through education and other shared contexts. For NNS students at the start of an undergraduate degree these shared primings may be different to those of NS students. In the case of the Chinese students in this study, initial primings are formed through input such as Intensive Reading classes with their focus on grammatical rules and translated

sentences, and the informal reading texts found in university exams. Using Hoey's theoretical framework, I have investigated characteristics of Chinese undergraduates' writing overall, across year groups, and across disciplines.

Below I summarize the findings for the research questions based on these areas:-

RQ 1: What are the distinguishing characteristics of writing in English in a corpus of Chinese undergraduates' assignments in the UK?

The first characteristic to be uncovered in the study is the significantly shorter length of the Chinese assignments and the significantly lower mean sentence length (MSL) of these texts. These findings can be partially accounted for by the greater use of visuals and lists in the Chinese texts, since words within tables and figures are deleted from the text files along with their captions, and writing in lists entails fewer words than writing in connected prose. Despite the lower MSL, however, the Chinese texts have a significantly higher mean word length (MWL) than the English texts, and I proposed that this may be due to the omission of short functional words (e.g. *the*) in the writing overall, and in particular in the abbreviated writing within lists. The overall profile of the Chinese texts is therefore somewhat different to the English texts, with the former containing less writing within prose sections of texts, and writing in shorter sentences with longer words.

The first category of keywords identified in the Chinese texts was that of informal language, though this was limited to a very small range of lexis (e.g. *lots (of), a bit of,* and the connectors *besides, what's more, last but not least*). However the English texts contain significantly higher usage of contracted verb forms (e.g. *doesn't, can't*), which Biber et al. (1999) class as informal language. Few occurrences of informal items in either corpus can be accounted for by personal, reflective sections of writing, and I hypothesized that student academic writing may simply have become less formal than was claimed by Biber et al. in 1999. The finding that both Chinese and English students make some use of language often described as informal contradicts findings in the literature which state that NNS writers make far greater use of this feature.

While the use of informal language in the datasets does not display the stark difference between NS and NNS writers claimed by much of the literature in the field, the findings on connector usage are in broad agreement with those of previous studies. The Chinese students make higher use of particular connectors such as *on the other hand* and *in the long run,* perhaps clinging to these familiar items as 'lexical teddy bears' (Hasselgren, 1994: 237). Also in line with the literature is the finding that the Chinese students prefer to employ connectors sentence-initially rather than in medial or sentence-end position.

The third keyword category discussed was that of first person pronouns, and my investigation revealed that the plural form is used significantly more by the Chinese student group while, in contrast, the English students make significantly greater use of the first person singular. These differences can be partly explained by the higher presence of reflective writing within the English corpus; for example, some of the English assignments are written in a formal essayist style for the main body of writing (with little use of *I*) and have an end section (labelled 'self-reflection task' or similar) which is written in a more personal, reflective style (resulting in end bursts of *I*). In comparison, the Chinese texts contain little reflective writing. The Chinese students' high use of *we* is partly accounted for by sections recounting methodological procedures (e.g. 'we used values of sample' (sic), 'we need to add anti-bumping granules'), whereas the English students use both *I* and *we* for this purpose.

A major finding of this study which has not previously been reported in the literature is the higher use of visuals and lists by the Chinese students, and this comprises the fourth keyword category which was discussed in Chapter 5. Keywords in the Chinese texts include chunks referring to visuals such as *the figure, according to the* and *illustrated in*; comparing the frequencies of tagged items in the corpora revealed that figures, tables, formulae and listlikes occur significantly more frequently within the Chinese students' assignments than the English ones. The use of these features could be viewed as compensatory strategies on the part of the Chinese students for coping with the extensive writing required at tertiary level

in the UK. Alternatively, and the position taken in this thesis, visuals and lists provide ways of presenting information clearly and concisely, both of which are highly-valued qualities of academic writing. Additionally, using visuals and lists is a more contemporary way of presenting information and may reflect the genres which students have been exposed to, namely websites and e-texts. Since both student groups have gained high marks, it is clear that visuals and lists are acceptable within the disciplines in the study.

RQ 2: Are there any variations in the characteristics identified in this study between years 1/2 and year 3?

The comparison of years 1/2 and year 3 revealed considerable variation in the characteristics previously identified. As might be predicted, year 3 texts for both Chinese and English cohorts are longer than those of the year 1/2 cohorts due to the increase in lengthy genres expected of final year undergraduates. Less predictable is the finding that the Chi3 assignments have a higher MSL and lower MWL than the Chi12 texts. This discrepancy is due in part to the significantly higher use of 1-letter 'words' such as the first person singular in Chi3. Additionally, the Chinese year 3 texts employ significantly more 1-letter items such as single character abbreviations (e.g. within descriptions of formulae), and numerals (often within lists) in their writing, both of which help to contribute to a lower MWL.

Most of the distinguishing characteristics of the Chinese students' writing found through keyword analysis are less apparent in year 3. The use of informal language, particular connectors, and the first person plural are less prevalent in Chi3 compared to Chi1/2, suggesting that the Chinese students in the Chi3 corpus have a broader range of linguistic resources and have adopted conventionalized writing features of the academy. This accords with Hoey's (2005) lexical priming theory in which individuals are primed by each new linguistic encounter: year 3 students in a UK university (of any L1) have had many more priming opportunities within academic writing than year 1/2 students. This possibility is supported by the analysis of extracts from Chinese ELT textbooks (reported in 6.4); these texts give an indication of Chinese students' previous exposure to 'academic writing' in

English, revealing this to consist of mixed formality levels and high use of pronouns and connectors.

In contrast, the use of visuals and listlikes is higher in the Chi3 texts than those of Chi1/2, as measured by counts of tagged features. It is difficult to ascertain from a corpus study why this might be the case, though I proposed in Chapter 6 that the increase is due to the year 3 Chinese students adopting these successful strategies in their writing. Since this is a quasi-longitudinal study rather than one following the same students over time, variation between Chi12 and Chi3 can be viewed as potential indicators of difference across the year groups.

The categorization of 4-grams in Chapter 6 revealed commonalities across the student corpora, with similar numbers of verb-based 4-grams in both Chinese and English corpora from all year groups. This similarity across student groups conflicts with the research literature which consistently describes NNS writing as more 'speech-like' than NS writing, (i.e. using more chunks containing verbs).

RQ 3: In what ways do disciplines affect the identified characteristics of Chinese undergraduate writing in English?

Three disciplines were investigated to answer this research question: Biology, Economics and Engineering, with variations found both across discipline groups and also between Chinese and English student groups within a discipline. Among the variations found between the three disciplines overall were the high use of numerals in both Biology and Engineering, and the use of fixed connectors pertaining to prediction in Economics (e.g. *in the long run*). In terms of pronoun use, for both student groups Economics texts use first person pronouns the most, Engineering texts less, and Biology texts use them the least. Finally, both student groups made highest use of listlikes in Engineering and lowest in Biology.

Of high statistical significance between student groups within the same discipline is the use of *I* in Eng-Engineering compared to *I* in Chi-Engineering, and I suggested that this may be due to the higher incidence of reflective writing in the English Engineering students' writing.

Additionally, the Chinese students' use of tables, figures and listlikes was significantly higher in some disciplines than the English students' use in those disciplines. Detailed analysis of pairs of assignments within the same discipline, module and topic, by Chinese and English writers, suggested that the use of visuals and lists function as different, yet equally valued, ways of writing. Data from interviews with BAWE lecturers adds support to this, indicating that concision and the use of visuals are valued features of student writing.

8.2 Implications

In this section, I discuss some implications arising from the study in relation to pedagogy, the methodology of Corpus Linguistics and to studies of learner corpora.

Pedagogical implications

Although pedagogy is not a major focus of this thesis, I discuss three implications of the findings for the teaching of academic writing. First, undergraduate writing in UK universities presents additional challenges to this student group since it is very different from Chinese students' previous experience of academic writing (i.e. textbooks with lists of connectors, Intensive Reading lessons with a focus on word-by-word translation). While it may not be possible in the PRC for Chinese students who study abroad to be more effectively prepared through secondary school textbooks and teaching, it may be more possible for university foundation classes in the UK to consider the kinds of writing students will undertake at undergraduate level. In EAP classes, access to a corpus of student academic writing (e.g. BAWE, MICUSP) and the means to specify particular disciplines and genres permits analysis of the types of assignment that students are required to write and assists tutors in understanding how different disciplines assess writing.

Second, I consider the implications of the finding that successful Chinese students' assignments frequently use visuals in their writing. Johns (1998: 183) points out that most applied linguists (and by implication most EAP tutors) are 'trained in the humanities, where words are central to disciplinary values and argumentation'. In writing centres and EAP

classrooms tutors may 'find themselves relying on disciplinary norms they are familiar with' (Gardner and Holmes, 2009: 251); these norms are likely to include a concentration on 'linear text' (Johns, 1988: 183) rather than on the interaction of visuals with text. This privileging of continuous prose over the use of graphs, diagrams, and images disadvantages not only those students who need to acquire competence in the production and comprehension of visuals in disciplines such as Biology and Economics, but also those who may be more visually-oriented or who may find it preferable to provide part of their response through graphical means. Interpreting visuals is becoming an increasingly important life skill with the advent of Web 2.0, as students are required to interpret data not only in their academic disciplines but also in newspapers, magazines, and in personal spheres such as financial information (Sorapure, 2010). Kress (2009: 57) points out that the 'semiotic reach' of modes, defined as 'what can be expressed readily or at all' by modes such as image, writing, layout, may vary according to culture; thus, what is achieved by writing in one culture may be achieved through image in another. It may also be the case that Chinese students more readily employ the visual mode, to extend or support that of language (e.g. when explaining biological phenomena).

Methodological implications

In my exploration of corpus linguistic procedures for n-gram type and token counting (4.3.3), I discussed the range of values used in parameters in empirical studies reported in the literature. In addition to choosing the minimum frequency threshold and minimum text dispersion thresholds, criteria such as tokens occurring in texts from a minimum number of individuals or other groupings (e.g. disciplines, genres) are set for each study. While the lack of consensus in parameter-setting allows researchers to experiment in order to retrieve the desired quantity of data, it renders the replicability of studies a more difficult feat. One way in which this can be helped is through greater transparency in the both description of datasets and in the procedures carried out on the data (e.g. giving raw figures as well as normalized ones, and providing details of the calculations used). I also argued in Chapter 4 that comparing the number of n-gram types across differently-sized corpora is fraught with difficulties. When I attempted to follow Biber and Barbieri's (2007) method for the

comparative normalization of types I achieved varying results each time, leading me to abandon this procedure. A more consistent procedure is needed to allow comparisons across different sizes of corpora.

A further implication for methodology relates to the difficulty I experienced in applying Hyland's (2008a,b) functional classification of 4-grams to authentic data. Hyland's well-exemplified categories initially appear straightforward, but in practice it is frequently difficult to place a 4-gram type into a single functional category. The (perhaps subjective) placements of common n-gram types then influence the overall findings. Again, this difficulty points to the importance of transparency of method, and the value of questioning (rather than accepting) the procedures followed in previous studies.

Implications for studies of NNS corpora

Throughout this thesis, I have argued that findings from research using learner corpora (e.g. ICLE) cannot be unquestioningly applied beyond the genre of very short argumentative essays. Learner corpus essays comprising 500 words are written with different goals and motivations to both 6000-word-plus professional academic articles and 2000-3000 word undergraduate assignments. The use of informal language, first person pronouns, connectors, and the presence or absence of visuals are among those features which differ according to genre, audience and purpose. Findings from learner corpora can inform us how certain NNS cohorts write short pieces of unassessed writing or short essays in a test situation, but claims cannot be made for learner writing beyond these situations. In this study, I have taken great care to compile a corpus of Chinese undergraduate texts with full transparency as to the contents, provenance and restrictions of the corpus. While some findings from learner corpora have been confirmed in this study (e.g. use of fixed connectors by Chinese students), others have not (e.g. high use of informal language by Chinese students), and these differences highlight the importance of avoiding over-generalization of findings to different genres of writing.

8.3 Limitations and suggestions for future research

The limitations given in this section are accompanied by suggestions for how they could be overcome in future studies.

The corpora in this study

This study contains assignments predominantly from BAWE, and the use of this existing corpus has both positive and negative implications. Assignments in the BAWE project were collected from just four universities (Oxford Brookes, Reading, Warwick and Coventry), and texts in each discipline were predominantly collected from a single university. There is thus a risk that most assignments in a discipline may be from a single university department, rendering them less justifiably representative of all UK student writing. However, this restriction in the collection strategy has the advantage of making it possible to compare NS and NNS students' assignments from the same course, module, and topic (as reported in Chapter 7). While I was not able to interview the student writers themselves, the BAWE project is supported by lecturer interviews in three of the universities and this provided useful insights into what is valued within student writing (Nesi and Gardner, 2006; Leedham, 2009). With the additionally-collected texts, the resulting corpus of 280,000 words in this study is the largest and most homogenous collection of UK university undergraduate assignments from Chinese students currently available.

This study examined Chinese and English students' writing only, and it is unclear whether the findings can be generalized beyond these groups. Future research could employ corpora of assignments from students with other L1s, to replicate the study (e.g. the VESPA study underway at Louvain). While my study focused on proficient student writing it might also be fruitful to compare these proficient undergraduate assignments with ones receiving a lower score (i.e. Third or Pass level in the UK). However, it could be difficult to collect assignments with lower grades, as students may be unwilling to offer these for research. In collecting additional assignments, I received no assignments with a mark below 50% even though no

proficiency level was given in the requirements. A study of low-scoring assignments could be combined with tutor and student interviews to ascertain the rationale for the score.

The methodology of Corpus Linguistics

The creation of a corpus by definition involves stripping a text from its surroundings and inevitably context is lost. One means of overcoming the loss of context in the case of written data is to combine Corpus Linguistics with detailed reading of a sample of texts to ground the research and maintain a link with the texts' origins, as carried out in the reading of paired assignments in Chapter 7. A further risk in Corpus Linguistics is assuming that the use of a large amount of data means that a corpus is fully representative of a particular group of language users. As Hoey (2005: 14) argues, a corpus 'represents no-one's experience of the language' but is a collection of primings, producing merely an approximate indicator of language use. All that a corpus can do, then, is to 'indicate that certain primings are likely to be shared by a large number of speakers, and only in that sense is priming independent of the individual' (Hoey, 2005: 15).

Corpus Linguistics is often assumed to be an objective investigative means whereby the human researcher can recede into the background. However, except in a fully-automated procedure such as concgram extraction, human intuition has a major role to play, from deciding on the constitution of the corpus, to setting the parameters for language analysis, and finally interpreting the results. Moreover, analyzing the most frequent n-grams in a corpus, whatever search parameters are used, by definition ignores lesser-used n-grams. A fundamental problem in making statements as to the structural and functional nature of, for example, 'top-sliced' 4-grams is that it awards attention only to those multiply-used, fixed 4-grams (such as *on the other hand* or *as well as the*), and not to the large numbers of 4-grams with similar meanings but with internal variability (e.g. *it is more important* and *it is very important*) or with a different outward form (e.g. *it is important* and *it is significant that*). It is useful, then, to consider n-grams with a common frame or with a common grammatical item, as is the case with semantic sequences (Hunston, 2008). These 'sequences of

meaning elements' (Hunston, 2008) are identifiable by individuals but cannot be established through n-gram analysis as the outward forms may differ (see also discussion in 3.3).

The corpus linguistic procedures carried out also have limitations in investigations of multimodal text features such as the use of visuals and lists and the layout of assignments. BAWE tagging of visuals marks the position of (deleted) visuals and captions within a text (though lists and listlikes are tagged but remain in the text). This tagging affects the wordcounts of assignments, since words within tables, figures and captions (which may be whole paragraphs) are omitted, removing significant stretches of data. More extensive tagging of visuals would enable the quantification of this semiotic resource; however, any automatic tagging entails the use of pre-determined categories of features within the data and moves away from a corpus-driven approach. To explore the ways in which students use non-linguistic semiotic resources, I argue that a whole text analysis is necessary in order to examine the context and the way in which visuals are employed within the assignment. Exploration of visual modes is pertinent as meaning is becoming increasingly mediated through the visual. Reading and analysing whole texts keeps the individual assignment at the forefront; in corpus linguistic research there is a risk that a sense of the whole will be lost in the focus on detailed patterns across a dataset.

The privileging of the mode of writing within corpus linguistic analysis ignores the impact of images on the reader and marker of the assignment. While it is possible to tag visual features and indicate the nature of each feature (e.g. length, position on the page, use of colour), this relies on pre-established categories, and cannot achieve the detail enabled by qualitative analysis of assignments. As an outsider to both the process of marking these particular assignments and to the specific disciplines, I am not viewing the assignments in the same way as either the student/writer or the lecturer/marker. Reading PDF versions of the assignments, while allowing more contextual information than is permitted in the corpus, permits only the 'flat' digital product to be analyzed. Aspects of the intended version may be missing in the digital reproduction (e.g. additional artefacts, aspects of presentation such as a colour binding). An additional layer of complication is that an assignment may be submitted

electronically to the tutor, then a hard copy printed and marked, resulting in the omission of the mode of colour and of the interactive element of live weblinks (thus leading to loss or change in the text's cohesion) (cf. Wyatt-Smith and Kimber, 2009). Published academic writing in the form of research articles and textbooks contains more visuals and less prose than is often required in student writing. In this sense, the Chinese students in the study are writing more in the style of professional writers, though their motivations may be different.

In addition to a text-bound study, the focus on visuals and lists in particular could be extended through interviews with both lecturers and students to investigate the insider perspective of the use of these features. This would answer questions such as the extent to which lecturers value students' use of visuals and lists, how these views vary across disciplines, and how aware students themselves are in their use of these features throughout their undergraduate studies. Most previous work on student writing has focused on either foundation level (in the case of NNSs) or postgraduate level writing (for both NSs and NNSs) and much more remains to be done regarding the exploration of undergraduate assessed writing. More studies of undergraduate assignments from both NSs and NNS writers are needed, since both groups are likely to experience difficulties in meeting the challenge of writing within the academy and may accomplish this in varying ways.

References

Abasi, A. R., & Akbari, N. (2008). Are we encouraging patchwriting? Reconsidering the role of the pedagogical context in ESL student writers' transgressive intertextuality. *English for Specific Purposes, 27*(3), 267-284.

Adamson, B. (2004). *China's English: A History of English in Chinese Education*. Hong Kong: Hong Kong University Press.

Alexander, O. (2001). *In general and in particular: When to pay attention to detail in text*. Paper presented at the BALEAP PIM conference: Focus on Chinese learners. Retrieved 20/01/2011, from http://www.baleap.org.uk/pimreports/2001/shu/abstracts.htm#secondary

Altenberg, B. (1998). On the phraseology of spoken English: The evidence of recurrent word-combinations. In A. P. Cowie (Ed.), *Phraseology* (pp. 101-122). Oxford: Oxford University Press.

Altenberg, B., & Tapper, M. (1998). The use of adverbial connectors in advanced Swedish learners' written English. In S. Granger (Ed.), *Learner English on Computer* (pp. 80-93). London/New York: Addison Wesley Longman.

Anthony, L. (2008). Antconc. Retrieved 20/02/2011, from http://www.antlab.sci.waseda.ac.jp/software.html

Archer, A. (2006). A multimodal approach to academic 'literacies': Problematising the visual/verbal divide. *Language and Education, 20*(6), 449-462.

Baba, K. (2009). Aspects of lexical proficiency in writing summaries in a foreign language. *Journal of Second Language Writing, 18*(3), 191-208.

Baigent, M. (2005). Multi-word chunks in oral tasks. In J. R. Willis & C. Edwards (Eds.), *Teachers Exploring Tasks* (pp. 157-170). Basingstoke: Palgrave Macmillan.

Baker, P. (2004). Querying keywords: Questions of difference, frequency, and sense in keywords analysis. *Journal of English Linguistics, 32*(4), 346-359.

Bakhtin, M. M. (1981). *The Dialogic Imagination: Four Essays* (C. Emerson & M. Holquist, Trans.). Austin: University of Texas Press.

Banerjee, J., Franceschina, F., & Smith, A. M. (2007). *Documenting features of written language production typical at different IELTS band score levels*: (IELTS Funded Research Project, Round 10, 2004).

Bassetti, B. (2005). Effects of writing systems on second language awareness: Word awareness of English learners of Chinese as a foreign language. In V. Cook & B. Bassetti (Eds.), *Second Language Writing Systems* (pp. 335-356). Clevedon: Multilingual Matters.

Bassetti, B. (2009). Effects of adding interword spacing on Chinese reading: A comparison of Chinese native readers and English readers of Chinese as a second language. *Applied Psycholinguistics, 30*, 757-775.

Bazerman, C. (2001). Distanced and refined selves: Educational tensions in writing with the power of knowledge. In M. Hewings (Ed.), *Academic Writing in Context: Implications and Applications* (pp. 23-29). Birmingham: Birmingham University Press.

Becher, T. (1989). *Academic Tribes and Territories*. Milton Keynes: The Society for Research into Higher Education/Open University Press.

Becher, T. (2004). The significance of disciplinary differences. *Studies in Higher Education, 19*(2), 151-161.

Becker, J. D. (1975). The phrasal lexicon. In R. C. Schank & B. L. Webber (Eds.), *Theoretical Issues in Natural Language Processing: 1* (pp. 70-73). Cambridge, MA.

Berber Sardinha, A. P. (2004). *Lingüística de Córpus*. Sao Paulo, Brazil: Manole.

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes, 26*(3), 263-286.

Biber, D., & Conrad, S. (1999). Lexical bundles in conversation and academic prose. In H. Hasselgard & S. Oksefjell (Eds.), *Out of Corpora: Studies in Honour of Stig Johansson* (pp. 181-190). Amsterdam: Rodopi.

Biber, D., Conrad, S., & Cortes, V. (2004). 'If you look at...': Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*(3), 371-405.

Biber, D., Conrad, S., Reppen, R., Byrd, P., & Helt, M. (2003). The authors respond: Strengths and goals of multidimensional analysis. *TESOL Quarterly, 37*, 151-155.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Harlow, Essex: Pearson.

Biber, D., Johansson, S., Reppen, R., Byrd, P., & Helt, M. (2002). Speaking and writing in the university: A multidimensional comparison. *TESOL Quarterly, 36*, 9-48.

Biglan, A. (1973). The characteristics of subject matter in different academic areas. *Journal of Applied Psychology, 57*(3), 195-203.

Bloor, M., & Bloor, T. (2001). There'll be some changes made: Predicting future events in academic and business genres. In M. Hewings (Ed.), *Academic Writing in Context* (pp. 182-198). Birmingham: The University of Birmingham Press.

Bolton, K. (2008). English in Asia, Asian Englishes, and the issue of proficiency. *English Today, 24*(2), 3-12.

Bolton, K., Nelson, G., & Hung, J. (2002). A corpus-based study of connectors in student writing: Research from the International Corpus of English in Hong Kong (ICE-HK). *International Journal of Corpus Linguistics, 7*(2), 165-182.

Boyle, J. (2000). Education for teachers of English in China. *Journal of Education for Teaching, 26*(2), 147-155.

Braine, G., & McNaught, C. (2007). Adaptation of the 'writing across the curriculum' model to the Hong Kong context. In J. Liu (Ed.), *English Language Teaching in China*. London: Continuum.

British Council. (2010a). China Market Introduction. Retrieved 17/08/2010, from http://www.britishcouncil.org/eumd-information-background-china.htm

British Council. (2010b). Education UK partnership: China country profile. Retrieved 21/3/2011, from http://www.britishcouncil.org/eumd-information-profiles-partnership.htm

British Embassy in Beijing. (2010). Progressive partnerships. *UK in China.* Retrieved 20/04/2010, from http://ukinchina.fco.gov.uk/en/news/?view=PressR&id=32930682

Bruce, I. (2010). Textual and discoursal resources used in the essay genre in Sociology and English. *Journal of English for Academic Purposes, 9*(3), 153-166.

Burnard, L. (2007). Reference Guide for the British National Corpus. XML. Retrieved 14/07/2009, from http://www.natcorp.ox.ac.uk/XMLedition/URG/

Bygate, M., Skehan, P., & Swain, M. (2001). *Researching Pedagogic Tasks: Second Language Learning, Teaching and Testing*. London: Longman.

Cai, G. (2011). The tertiary English language curriculum in China and its delivery: A critical study. Unpublished PhD. The Open University.

Canale, M., & Swain, H. (1980). Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing. *Applied Linguistics, 1*, 1-47.

Chambers, E., & Northedge, A. (1997). *The Arts Good Study Guide*. Milton Keynes: Open University Press.

Channell, J. (1994). *Vague Language*. Oxford: Oxford University Press.

Charles, M. (2003). `This mystery...': A corpus-based study of the use of nouns to construct stance in theses from two contrasting disciplines. *Journal of English for Academic Purposes, 2*(4), 313-326.

Charles, M. (2006). Phraseological patterns in reporting clauses used in citation: A corpus-based study of theses in two disciplines. *English for Specific Purposes, 25*(3), 310-331.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4*, 55-81.

Chen, Y. (2009). Investigating lexical bundles across learner writing development. Unpublished PhD. Lancaster University.

Chen, Y., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning and Technology, 14*(2), 30-49.

Cheng, W., Greaves, C., Sinclair, J. M., & Warren, M. (2009). Uncovering the extent of the phraseological tendency: Towards a systematic analysis of concgrams. *Applied Linguistics, 30*(2), 236-252.

Cheng, W., Greaves, C., & Warren, M. (2006). From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics, 11*, 411-433.

Cheng, X. (2000). Asian students' reticence revisited. *System, 28*(3), 435-446.

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.

Chuang, F., & Nesi, H. (2006). An analysis of formal errors in a corpus of L2 English produced by Chinese students. *Corpora, 1*(2), 251-271.

Clark, M. (2003). Computer Science: A hard-applied discipline? *Teaching in Higher Education, 8*, 71-87.

Clark, R., & Gieve, S. (2006). On the discursive construction of 'The Chinese Learner'. *Language, Culture & Curriculum, 19*(1), 54-73.

Cobb, T. (2003). Teaching and researching writing. *System, 31*(1), 132-136.

Conklin, K., & Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and non native speakers? *Applied Linguistics, 29*(1), 72-89.

Cook, G. (1998). The uses of reality: A reply to Ronald Carter. *English Language Teaching Journal, 52*(1), 57-63.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from History and Biology. *English for Specific Purposes, 23*(4), 397-423.

Cortes, V. (2006). Teaching lexical bundles in the disciplines: An example from a writing intensive History class. *Linguistics and Education, 17*(4), 391-406.

Craft, A. (2001). Neuro-linguistic programming and learning theory. *Curriculum Journal, 12*(1), 125-136.

Crème, P., & Lea, M. R. (2003). *Writing at University: a Guide for Students* (2nd ed.). Buckingham: Open University Press

Crick, F. (1979). Thinking about the brain. *Scientific American, 9*, 218-232.

Croft, W., & Cruse, D. A. (2004). Cognitive Linguistics. *Cambridge, Cambridge University Press*.

Cross, J., & Hitchcock, R. (2007). Chinese students' (or students from China's) views of UK HE: Differences, difficulties and benefits, and suggestions for facilitating transition. *The East Asian Learner, 3*(2), 1-31.

Cross, J., & Papp, S. (2008). Creativity in the use of verb + noun combinations by Chinese learners of English. In G. Gilquin, S. Papp & M. B. Díez-Bedmar (Eds.), *Linking up Contrastive and Learner Corpus Research* (pp. 57-81). Amsterdam/Atlanta: Rodopi.

Culpeper, J. (2009). Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's Romeo and Juliet. *International Journal of Corpus Linguistics, 14*, 29-59.

Danielsson, P. (2008). WordCount2. Birmingham: University of Birmingham.

De Cock S. (2000). Repetitive phrasal chunkiness and advanced EFL speech and writing. In Mair C. and Hundt M. (eds) Corpus Linguistics and Linguistic Theory. *Papers from the Twentieth International Conference on English Language Research on Computerized Corpora (ICAME 20), Freiburg im Breisgau 1999*. Amsterdam: Rodopi, 51-68.

De Cock, S., & Granger, S. (2004). *High frequency words: The bête noire of lexicographers and learners alike. A close look at the verb 'make' in five monolingual learners dictionaries of English.* Paper presented at the Eleventh EURALEX International Congress, Université de Bretagne-Sud: Lorient.

Dinolfo, J., Heifferon, B., & Temesvari, L. A. (2007). Seeing cells: Teaching the visual/verbal rhetoric of Biology. *Journal of Technical Writing and Communication, 37*(4), 395-417.

Dongping, Y. (2006). Pursuing harmony and fairness in education. *Chinese Education and Society, 39*(6), 3-44.

Douglas, S. R. (2010). Non-native English speaking students at university: Lexical richness and academic success. Unpublished PhD. University of Calgary.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics, 19*(1), 61-74.

Durkin, K. (2010). Adapting to western norms of critical argumentation and debate. In M. Cortazzi & L. Jin (Eds.), *Researching Chinese Learners: Skills, Perceptions and Intercultural Adaptation*. Basingstoke: Palgrave Macmillan.

Durkin, K., & Main, A. (2002). Discipline-based study skills support for first-year undergraduate students. *Active Learning in Higher Education, 3*(1), 24-39.

Durrant, P. (2008). High-frequency collocations and second language learning. Unpublished PhD. Nottingham University.

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes, 28*(3), 157-169.

Dzau, Y. F. (1990). *English in China*. Hong Kong: API Press.

Ebeling, S. O., & Heuboeck, A. (2007). Encoding document information in a corpus of student writing: The British Academic Written English corpus. *Corpora, 2*(2), 241-256.

Ellis, N. (1997). Vocabulary acquisition: Word structure, collocation, word-class, and meaning. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, Acquisition and Pedagogy* (pp. 122-139). Cambridge: Cambridge University Press.

Ellis, N. C., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, Corpus Linguistics, and TESOL. *TESOL Quarterly, 42*, 375-396.

Ellison, N. B., Steinfield, C., & Lampe, C. (2007). The benefits of Facebook 'friends': Social capital and college students' use of online social network sites. *Journal of Computer Mediated Communication, 12*(4), 11-43.

Elton, L. (2010). Academic writing and tacit knowledge. *Teaching in Higher Education, 15*(2), 151-160.

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing, 4*(2), 139-155.

Erman, B. (2007). Cognitive processes as evidence of the idiom principle. *International Journal of Corpus Linguistics, 12*, 25-53.

Field, Y., & Yip, L. (1992). A comparison of internal cohesive conjunction in the English essay writing of Cantonese speakers and native speakers of English. *RELC Journal, 23*(1), 15-28.

Fillmore, C., Kay, P., & O'Connor, M. C. (1988). Regularity and idiomaticity in grammatical constructions. *Language, 64*(501-38).

Flowerdew, L. (2003). A combined corpus and systemic-functional analysis of the problem-solution pattern in a student and professional corpus of technical writing. *TESOL Quarterly, 37*(3), 489-511.

Foster, P. (2001). Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. In M. Bygate, P. Skehan & M. Swain (Eds.), *Task-Based Learning: Language Teaching, Learning and Assessment* (pp. 75-97). London: Longman.

Foster, P., Tonkyn, A., & Wrigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics, 21*(3), 354-375.

Fox, J., & Curtis, A. (2010). IELTS: International English Language Testing System. In L. Cheng & A. Curtis (Eds.), *English Language Assessment and the Chinese Learner*. pp.112-120. New York: Routledge.

Galloway, I. (2005). Computer learner corpora and their pedagogical application. *TESOL Quarterly, 39*, 333-340.

Gan, Z., Humphreys, G., & Hamp-Lyons, L. (2004). Understanding successful and unsuccessful EFL students in Chinese universities. *Modern Language Journal, 88*, (2), pp. 229-244.

Ganobcsik-Williams, L. (2004). A report on the teaching of academic writing in UK Higher Education. *London: Royal Literary Fund*.

Gao, M. (2000). *Mandarin Chinese*. Oxford: Oxford University Press.

Gardner, S. (2008). Integrating ethnographic, multidimensional, corpus linguistic and systemic functional approaches to genre description: An illustration through university History and Engineering assignments. In Steiner, E and Neumann, S. (Eds.) *ESFLCW 2007: Data and Interpretation in linguistic analysis. Proceedings of the 19th European Systemic Functional Linguistics Conference and Workshop* 23rd - 25th July 2007, Saarbrücken, Germany: Universität des Saarlandes (pp 1-34). http://scidok.sulb.uni-saarland.de/sulb/portal/esflcw/

Gardner, S., & Holmes, J. (2009). Can I use headings in my essay? Section headings, macrostructures and genre families in the BAWE corpus of student writing. In M. Charles, S. Hunston & D. Pecorari (Eds.), *At the Interface of Corpus and Discourse: Analysing Academic Discourses* (pp. 251-271). London: Continuum.

Gardner, S., & Nesi, H. (2008). *A new categorisation of university student writing tasks*. Paper presented at: Language Issues in English-Medium Universities: A Global Concern, University of Hong Kong, Hong Kong.

Gerbic, P. (2005). *Chinese learners and computer mediated communication: Balancing culture, technology and practice.* Paper presented at the ASCILITE Conference: Balance, Fidelity, Mobility: Maintaining the Momentum?, Brisbane, Australia.

Gibbs, G. (2006). Why assessment is changing. In C. Bryan & K. Clegg (Eds.), *Innovative Assessment in Higher Education* (pp. 11-22). New York: Routledge.

Gieve, S., & Clark, R. (2005). The Chinese approach to learning: Cultural trait or situated response? *System, 33*(2), 261-276.

Gilquin, G., & Paquot, M. (2007). *Spoken features in learner academic writing: Identification, explanation and solution.* Paper presented at the Fourth Corpus Linguistics Conference, Birmingham.

Gilquin, G., & Paquot, M. (2008). Too chatty: Learner academic writing and register variation. *English Text Construction, 1*, 41-61.

Gourlay, L. (2009). Threshold practices: Becoming a student through academic literacies. *London Review of Education, 7*(2), 181-192.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: collocations and formulae. In A. P. Cowie (Ed.), *Phraseology: Theory, Analysis and Applications* (pp. 145-160). Oxford: Clarendon Press.

Granger, S. (2002). A bird's-eye view of computer learner corpus research. In S. Granger, J. Hung & S. Petch-Tyson (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam/Philadelphia: Benjamins.

Granger, S. (2004). Computer learner corpus research: Current status and future prospects. *Language and Computers, 52*, 123-145.

Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (2009). International Corpus of Learner English v.2 Retrieved 15/05/2011, from http://www.uclouvain.be/en-cecl.html.

Granger, S., & Paquot, M. (2008). Disentangling the phraseological web. In S. Granger & F. Meunier (Eds.), *Phraseology* (pp. 27-49). Amsterdam: John Benjamins.

Granger, S., & Rayson, P. (1998). Automatic lexical profiling of learner texts. In S. Granger (Ed.), *Learner English on Computer* (pp. 119-131). London/New York: Addison Wesley Longman.

Grant, L., & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing, 9*(2), 123-145.

Greaves, C. (2009). ConcGram 1.0: A phraseological search engine. Amsterdam: John Benjamins.

Gries, S. T. (2008). Phraseology and linguistic theory: A brief survey. In S. Granger & F. Meunier (Eds.), *Phraseology* (pp. 3-25). Amersterdam: John Benjamins.

Groom, N. (2005). Pattern and meaning across genres and disciplines: An exploratory study. *Journal of English for Academic Purposes, 4*(3), 257-277.

Groom, N. (2007). *A corpus-driven study of phraseology in written academic English across two genres and two disciplines*. Paper presented at a Postgraduate Student Seminar, Birmingham: Birmingham University.

Gu, P. Y. (2003). Fine brush and freehand: The vocabulary-learning art of two successful Chinese EFL learners. *TESOL Quarterly, 37*(1), 73-104.

Gu, Q., & Brooks, J. (2008). Beyond the accusation of plagiarism. *System, 36*, 337-352.

Gu, Q., & Schweisfurth, M. (2006). Who adapts? Beyond cultural models of 'the' Chinese learner. *Language, Culture & Curriculum, 19*(1), 74-89.

Gui, S., & Yang, H. (2003). Chinese Learner English Corpus (CLEC). Retrieved 16/2/2011, from http://langbank.engl.polyu.edu.hk/corpus/clec.html

Guo. (2006). Verbs in the written English of Chinese learners: A corpus-based comparison between non-native speakers and native speakers. Unpublished PhD. University of Birmingham.

Halliday, M. A. K. (1994). *An Introduction to Functional Grammar* (2nd ed.). London: Edward Arnold.

Halliday, M. A. K., & Matthiessen, C. M. M. (2004). *An Introduction to Functional Grammar* (3rd ed.). London: Arnold.

Harwood, N. (2003). Person markers and interpersonal metadiscourse in academic writing: A multidisciplinary corpus-based study of student and expert texts. Unpublished PhD. University of Kent at Canterbury.

Harwood, N. (2005). `Nowhere has anyone attempted... In this article I aim to do just that': A corpus-based study of self-promotional *I* and *we* in academic writing across four disciplines. *Journal of Pragmatics, 37*(8), 1207-1231.

Harwood, N. (2009). (In)appropriate personal pronoun use in Politial Science. *Written Communication, 23*(4), 424-450.

Harwood, N., & Hadley, G. (2004). Demystifying institutional practices: Critical pragmatism and the teaching of academic writing. *English for Specific Purposes, 23*(4), 355-377.

Hasselgren, A. (1994). Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics, 4*(2), 237-258

Heuboeck, A., Holmes, J., & Nesi, H. (2008). *The BAWE Corpus Manual*. Warwick: Warwick University.

Hewings, A. (1999). Disciplinary engagement in undergraduate writing: An investigation of clause-initial elements in Geography essays. Unpublished PhD. University of Birmingham.

Hewings, A. (2004). Developing discipline-specific writing: An analysis of undergraduate geography essays. In L. J. Ravelli & R. A. Ellis (Eds.), *Analysing Academic Writing: Contextual Frameworks* (pp. 131-152). London: Continuum.

Hewings, A., & Coffin, C. (2007). Writing in multi-party computer conferences and single authored assignments: Exploring the role of writer as thinker. *Journal of English for Academic Purposes, 6*(2), 126-142.

Hewings, A., & North, S. (2006). Emergent disciplinarity: A comparative study of Theme in undergraduate essays in geography and history of science. In R. Whittaker, A. McCabe & M. O'Donnell (Eds.), *Language and Literacy: Functional Approaches* (pp. 266–281). London: Continuum.

Higher Education Statistics Agency (HESA). (2010). HESA Students in Higher Education Institutions. Retrieved 11/2/2011, from http://www.hesa.ac.uk/index.php/content/view/1398/161/

Hinds, J. (1987). Reader versus writer responsibility: A new typology of language. In U. Connor & R. Kaplan (Eds.), *Writing Across Languages: Analysis of L2 text* (pp. 141-152). Reading, MA: Addison-Welsey.

Hinkel, E. (2002). *Second Language Writers' Text: Linguistic and Rhetorical Features*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Hinkel, E. (2003). Simplicity without elegance: Features of sentences in L1 and L2 academic texts. *TESOL Quarterly, 37*, 275-301.

Hinkel, E. (2005). Hedging, inflating and persuading. *Applied Language Learning, 15*(1-2), 29–53.

Hoey, M. (2005). *Lexical Priming*. New York: Routledge.

Hoey, M. (2009). Corpus-driven approaches to grammar. In U. Römer & R. Schulze (Eds.), *Exploring the Lexis-Grammar Interface* (pp. 33-47). Amsterdam/Philadephia: John Benjamins.

Holliday, A. (2005). *The Struggle to Teach English as an International Language*. Oxford: Oxford University Press.

Hopkins, M. (2006). Policies without planning?: The medium of instruction issue in Hong Kong. *Language and Education, 20*(4), 270-286.

Hopper, P. (1987). Emergent Grammar. *Berkeley Linguistics Society, 13*, 139-157.

Hopper, P. (1998). Emergent Grammar. In M. Tomasello (Ed.), *The New Psychology of Language* (Vol. 1, pp. 155-175). Mahwah, New Jersey: Lawrence Erlbaum.

Horowitz, D. M. (1986). What professors actually require: Academic tasks for the ESL classroom. *TESOL Quarterly, 20*(3), 445-462.

Howard, D. (1983). *Cognitive Psychology: Memory, Language, and Thought*. New York: MacMillan.

Howarth, P. (1998). The phraseology of learners' academic writing. In A. P. Cowie (Ed.), *Phraseology* (pp. 161-186). Oxford: Oxford University Press.

Hu, G. (2001). *The People's Republic of China country report: English Language Teaching in the People's Republic of China*. Singapore: Nanyang Technological University.

Hu, G. (2005). Contextual influences on instructional practices: A Chinese Case for an ecological approach to ELT. *TESOL Quarterly, 39*, 635-660.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Hunston, S. (2006). Phraseology and system: A contribution to the debate. In G. Thompson & S. Hunston (Eds.), *System and Corpus* (pp. 55-80). London: Equinox.

Hunston, S. (2008). Starting with the small words: Patterns, lexis and semantic sequences. *International Journal of Corpus Linguistics, 13*(3), 271-295.

Hunston, S., & Francis, G. (2000). *Pattern Grammar*. Amsterdam: John Benjamins.

Hyland, K. (2002). Authority and invisibility: Authorial identity in academic writing. *Journal of Pragmatics, 34*(8), 1091-1112.

Hyland, K. (2003). Review of 'Genre in the classroom: Multiple perspectives'. *English for Specific Purposes, 22*(2), 213-215.

Hyland, K. (2004). Disciplinary interactions: Metadiscourse in L2 postgraduate writing. *Journal of Second Language Writing, 13*(2), 133-151.

Hyland, K. (2005a). *Metadiscourse*. London: Continuum.

Hyland, K. (2005b). Representing readers in writing: Student and expert practices. *Linguistics and Education, 16*(4), 363-377.

Hyland, K. (2008a). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics, 18*(1), 41-62.

Hyland, K. (2008b). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27*(1), 4-21.

Hyland, K. (2009). Corpus informed discourse analysis: The case of disciplinary engagement. In M. Charles, S. Hunston & D. Pecorari (Eds.), *At the Interface of Corpus and Discourse: Analysing Academic Discourses* (pp. 110-128). London: Continuum.

Hyland, K., & Milton, J. (1997). Qualification and certainty in L1 and L2 students' writing. *Journal of Second Language Writing, 6*(2), 183-205.

Hyland, K., & Tse, P. (2005). Hooking the reader: A corpus study of evaluative that in abstracts. *English for Specific Purposes, 24*(2), 123-139.

Hyland, K., & Tse, P. (2007). Is there an 'Academic Vocabulary'? *TESOL Quarterly, 41*(2), 235-253.

Hymes, D. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269 - 285). Harmondsworth: Penguin.

Ishikawa, S. (2010). Corpus of English Essays written by Asian university Students (CEEAUS). Retrieved 15/2/2011, from http://language.sakura.ne.jp/s/ceeause.html

Jackendoff, R. (1988). Conceptual semantics. In P. Violi, U. Eco & M. Santambrogio (Eds.), *Meaning and Mental Representation* (pp. 81-97). Bloomington: Indiana University Press.

Jarvis, S., Grant, L., Bikowski, D., & Ferris, D. (2003). Exploring multiple profiles of highly rated learner compositions. *Journal of Second Language Writing, 12*(4), 377-403.

Jenkins, J. (2009). Exploring attitudes towards English as a lingua franca in the East Asian context. In K. Murata & J. Jenkins (Eds.), *Global Englishes in Asian Contexts: Current and Future Debates* (pp. 40-56). Basingstoke: Palgrave Macmillan.

Jewitt, C. (2009). *The Routledge Handbook of Mulimodal Analysis*. London: Routledge.

Jiang, N., & Nekrasova, T. M. (2007). The processing of formulaic sequences by second language speakers. *Modern Language Journal, 91*(3), 433-445.

Jin, L., & Cortazzi, M. (1998). Dimensions of dialogue: Large classes in China. *International Journal of Educational Research, 29*, 739-761.

Jin, L., & Cortazzi, M. (2002). English language teaching in China: A bridge to the future. *Asia-Pacific Journal of Education, 22*(2), 53-64.

Jin, L., & Cortazzi, M. (2006). Changing practices in Chinese cultures of learning. *Language, Culture & Curriculum, 19*(1), 5-20.

Johns, A. M. (1998). The visual and the verbal: A case study in macroeconomics. *English for Specific Purposes, 17*(2), 183-197.

Johns, A. M. (2002). *Genre in the Classroom: Multiple Perspectives*. Mahwah, NJ: Lawrence Earlbaum.

Johns, T. F. (1991). Should you be persuaded: Two examples of data-driven learning materials. *English Language Research Journal, 4*, 1-16.

Jung, C. K. (2011). Understanding undergraduate Engineering laboratory reports (UELRs). Unpublished PhD. University of Warwick.

Kachru, B. B. (1985). Standards, codification and sociolinguistic realism: The English language in the outer circle. In R. Quirk & H. G. Widdowson (Eds.), *English in the World: Teaching and Learning the Language and Literatures*. Cambridge: Cambridge University Press.

Kaldor, S., & Rochecouste, J. (2002). General academic writing and discipline specific academic writing. *Australian Review of Applied Linguistics, 25*(2), 29-47.

Katz, S. (1996). Distribution of common words and phrases in text and language modelling. *Natural Language Engineering, 2*(1), 15-59.

Kennedy, C., & Thorp, D. (2007). A corpus investigation of linguistic responses to an IELTS Academic Writing task. In L. Taylor & P. Falvey (Eds.), *IELTS Collected Papers: Research in Speaking and Writing Assessment* (pp. 316–378). Cambridge: Cambridge University Press.

Kennedy, P. (2002). Learning cultures and learning styles: Myth-understandings about adult (Hong Kong) Chinese learners. *International Journal of Lifelong Education, 21*(5), 430-445.

Kinzley, S. (2011). The impact of a university pre-sessional course on the academic writing behaviours of a group of Chinese undergraduate students studying for a degree in media and cultural studies. Unpublished PhD. Lancaster University.

Kirkpatrick, A. (2004). Some thoughts on the Chinese learner and the teaching of writing *The East Asian Learner, 1*(1).1-15.

Kirkpatrick, A. (2006). *Teaching English across cultures. What do English language teachers need to know to know how to teach English?* Paper presented at the 19th Annual EA Education Conference, Perth, Australia.

Kolb, D. A. (1981). Learning styles and disciplinary differences. In A. W. Chickering (Ed.), *The Modern American College: Responding to the New Realities of Diverse Students and a Changing Society* (pp. 232-255). San Francisco: Jossey-Bass.

Kress, G. (2004). Reading images, multimodality, representation and new media. Retrieved 1/4/2011, from

http://www.knowledgepresentation.org/BuildingTheFuture/Kress2/Kress2Quicktime/Kress2Movie.html

Kress, G. (2009). What is mode? In C. Jewitt (Ed.), *The Routledge Handbook of Multimodal Analysis* (pp. 54-67). Abingdon: Routledge.

Kress, G., & Van Leeuwen, T. (1996). *Reading Images: The Grammar of Visual Design*. London: Routledge.

Kress, G., & Van Leeuwen, T. (2001). *Multimodal Discourse*. London: Arnold.

Langacker, R. (1990). Concept, Image, and Symbol: The Cognitive Basis of Grammar. *Berlin, Mouton de Gruyter*.

Lantolf, J. P., & Thorne, S. L. (2006). Sociocultural theory and second language acquisition. In B. Van Patten & J. Williams (Eds.), *Theories in Second Language Acquisition* (pp. 197-220). Mahwah, NJ: Erlbaum.

Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics, 27*, 590-619.

Lea, M. R. (2004). Academic literacies: A pedagogy for course design. *Studies in Higher Education, 29*(6), 739-756.

Lea, M. R., & Stierer, B. (2000a). Editor's introduction. In M. R. Lea & B. Stierer (Eds.), *Student Writing in Higher Education: New Contexts* (pp. 1-13). Buckingham: Open University Press.

Lea, M. R., & Stierer, B. (Eds.). (2000b). *Student Writing in Higher Education: New Contexts*. Buckingham: Society for Research into Higher Education/Open University Press.

Lea, M. R., & Street, B. (1998). Student writing in higher education: An academic literacies approach. *Studies in Higher Education, 23*(2), 157-172.

Lea, M. R., & Street, B. V. (2006). The 'academic literacies' model: Theory and applications. *Theory into Practice, 45*(4), 368-377.

Lee, D., & Chen, S. X. (2009). Making a bigger deal of the smaller words: Function words and other key items in research writing by Chinese learners. *Journal of Second Language Writing, 18*, 181-196.

Leedham, M. (2006). 'Do I speak better?' A longitudinal study of lexical chunking in the spoken language of two Japanese students. *The East Asian Learner, 2*(2), 1-15.

Leedham, M. (2009). From traditional essay to 'Ready Steady Cook' presentation: Reasons for innovative changes in assignments. *Active Learning in Higher Education, 10*(2), 191-206.

Lei, X. (2009). Understanding writing strategy use from a sociolcultural perspective: A multiple case study of Chinese EFL learners of different writing abilities Unpublished PhD. The University of Hong Kong.

Leki, I., & Carson, J. (1994). Students' perceptions of EAP writing instructions and writing needs across the disciplines. *TESOL Quarterly, 28*, 81-101.

Leung, C., Harris, R., & Rampton, B. (1997). The idealised native speaker, reified ethnicities, and classroom realities. *TESOL Quarterly, 31*(3), 543-560.

Lewis, M. (2000). *Teaching Collocation*. Hove: Language Teaching Publications.

Lewis, M. (2002). *The Lexical Approach*. London: Thomson Heinle.

Li, J. (2009). Advanced Chinese L2 learners' formulaic langauge use and acquisition in academic writing. Unpublished PhD. University of Nottingham.

Li, J., & Schmitt, N. (2009). The acquisition of lexical phrases in academic writing: A longitudinal case study. *Journal of Second Language Writing, 18*, 85-102.

Li, T. (2010). A study of metadiscourse in English academic essays: Similarities and differences among Chinese undergraduates, 2+2 Chinese undergraduates and English native undergraduates. Unpublished PhD. Warwick University.

Liang, L. H. (2010). Chinese students suffer as university entrance exams get a grip. *The Guardian*. Retrieved 08.09/2010, from http://www.guardian.co.uk/education/2010/jun/28/chinese-university-entrance-exams

Lillis, T. (1997). New voices in academia? The regulative nature of academic writing conventions. *Language and Education, 11*(3), 182-199.

Lillis, T. (1999). Whose 'common sense'? Essayist literacy and the institutional practice of mystery. In C. Jones, J. Turner & B. Street (Eds.), *Students Writing in the University* (pp. 127-147). Amsterdam: John Benjamins.

Lillis, T. (2001). *Student Writing: Access, Regulation, Desire*. London: Routledge.

Lillis, T. (2003). Student writing as ‘academic literacies’: Drawing on Bakhtin to move from critique to design. *Language and Education, 17*(3), 192-207.

Lillis, T. (2006). Moving towards an 'Academic Literacies' pedagogy: Dialogues of participation. In L. Ganobcsik-Williams (Ed.), *Teaching Academic Writing in UK Higher Education*: Palgrave Macmillan.

Lillis, T., & Scott, M. (2008). Defining academic literacies research: Issues of epistemology, ideology and strategy. *Journal of Applied Linguistics, 4*(1), 5-32.

Lillis, T., & Turner, J. (2001). Student writing in Higher Education: Contemporary confusion, traditional concerns. *Teaching in Higher Education, 6*(1).

Littlewood, W. (2007). Communicative and task-based language teaching in East Asian classrooms. *Language Teaching, 40*(3), 243-249.

Lu, Y. (2002). Linguistic characteristics in Chinese learner English. In M. Tan (Ed.), *Corpus Studies in Language Education* (pp. 49-60). Bangkok: IELE Press.

Luxon, T., & Robinson, S. (2006). *Embedding study support for Chinese students in content courses: MNGT 121 a case study*. Portsmouth: Portsmouth University.

Luzón, M. J. (2009). The use of *we* in a learner corpus of reports written by EFL Engineering students. *Journal of English for Academic Purposes, 8*(3), 192-206.

Martinez, I. A. (2005). Native and non-native writers' use of first person pronouns in the different sections of biology research articles in English. *Journal of Second Language Writing, 14*(3), 174-190.

Matsuda, P. K., Canagarajah, A. S., Harklau, L., Hyland, K., & Warschauer, M. (2003). Changing currents in second language writing research: A colloquium. *Journal of Second Language Writing, 12*(2), 151-179.

Mauranen, A. (1994). Two discourse worlds: Study genres in Britain and Finland. *Finlance. A Finnish Journal of Applied Linguistics 13*, 1-40.

Mayor, B. (2006). Dialogic and hortatory features in the writing of Chinese candidates for the IELTS test. *Language, Culture & Curriculum, 19*(1), 104–121.

Mayor, B., Hewings, A., North, S., & Swann, J. (2007). A linguistic analysis of Chinese and Greek L1 scripts for IELTS Academic Writing Task 2. In L. Taylor & F. Falvey (Eds.), *IELTS Collected Papers: Research in Speaking and Writing Assessment. Studies in Language Testing* (Vol. 19, pp. 250-313): Cambridge University Press.

McArthur, T. (1998). *The English Languages*. Cambridge: Cambridge University Press.

McArthur, T. (2008). English as an Asian language. *English Today, 19*(2), 19-22.

McCrostie, J. (2008). Writer visibility in EFL learner academic writing: A corpus-based study. *ICAME, 32*, 97-114.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-Based Language Studies: An Advanced Resource Book*. London: Routledge.

McKenny, J. (2005). Content analysis of dogmatism compared with corpus analysis of epistemic stance in student essays. *Information Design Journal, 13*(1), 40-49.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review, 63*, 81-97.

Milton, J. (1999). Lexical thickets and electronic gateways: Making text accessible by novice writers. In C. N. Candlin & K. Hyland. (Eds.), *Writing: Texts, Processes and Practices* (pp. 221-243). London: Longman.

Milton, J. (2001). Elements of a written interlanguage: A computational and corpus-based study of institutional influences on the acquisition of English by Hong Kong Chinese students. *Research Reports, 2*. Retrieved from http://repository.ust.hk/dspace/bitstream/1783.1/1055/1/john.pdf

Milton, J., & Hyland, K. (1999). *Assertions in students' academic essays: A comparison of English NS and NNS student writers*. Paper presented at the conference Language analysis, description and pedagogy. Retrieved on 13/05/2011from http://repository.ust.hk/dspace/bitstream/1783.1/1045/1/MILHYL2.pdf

Moon, R. (1998a). *Fixed Expressions and Idioms in English*. Oxford: Clarendon Press.

Moon, R. (1998b). Frequencies and forms of phrasal lexemes in English. In A. P. Cowie (Ed.), *Phraseology* (pp. 79-100). Oxford: Oxford University Press.

Moore, T., & Morton, J. (2005). Dimensions of difference: A comparison of university writing and IELTS writing. *Journal of English for Academic Purposes, 4*, 43-66.

Mu, C., & Carrington, S. (2007). An investigation of three Chinese students' English writing strategies. *TESL-EJ, 11*(1), 1-22.

Myers, G. (2001). 'In my opinion': The place of personal views in undergraduate essays. In M. Hewings (Ed.), *Academic Writing in Context* (pp. 63-78). Birmingham: Birmingham University Press.

Nathan, P. (2010). A genre-based study of pedagogical business case reports. Unpublished PhD. University of Birmingham.

Nation, P. (2008). *Teaching Vocabulary: Strategies and Techniques*. Boston: Heinle Cengage Learning.

Nattinger, J., & DeCarrico, J. (1992). *Lexical Phrases and Language Teaching*. Oxford: Oxford University Press.

Nesi, H. (2008a). *BAWE: An introduction to a new resource.* Paper presented at the 8th Teaching and Language Corpora Conference, Lisbon, Portugal.

Nesi, H. (2008b). *Corpora and EAP.* Paper presented at the The 6th Languages for Specific Purposes International Seminar, Universiti Teknologi Malaysia, Johor Bahru, Malaysia.

Nesi, H. (2011). BAWE: An introduction to a new resource. In A. Frankenberg-Garcia, L. Flowerdew & G. Aston (Eds.), *New Trends in Corpora and Language Learning* (pp. 213-228). London: Continuum.

Nesi, H., & Basturkmen, H. (2006). Lexical bundles and discourse signalling in academic lectures. *International Journal of Corpus Linguistics, 11*(3), 283-304.

Nesi, H., & Gardner, S. (2006). Variation in disciplinary culture: University tutors' views on assessed writing tasks. In R. Kiely, G. Clibbon, P. Rea-Dickins & H. Woodfield (Eds.), *Language, Culture and Identity in Applied Linguistics* (Vol. British Studies in Applied Linguistics, pp. 99-107). London: Equinox Publishing.

Nesi, H., & Gardner, S. (forthcoming, 2011). Chapter 7: The role of reflection. In H. Nesi & S. Gardner (Eds.), *Genres across the Disciplines: Student Writing in Higher Education*. Cambridge: Cambridge University Press.

Nesi, H., Gardner, S., Forsyth, R., Hindle, D., Wickens, P., Ebeling, S., et al. (2005). *Towards the compilation of a corpus of assessed student writing: An account of work in progress.* Paper presented at the Corpus Linguistics conference, Birmingham.

Nesi, H., Sharpling, G., & Ganobcsik-Williams, L. (2004). Student papers across the curriculum: Designing and developing a corpus of British student writing. *Computers and Composition, 21*, 439-450.

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics, 24*(2), 223-242.

Neumann, R., Parry, S., & Becher, T. (2002). Teaching and learning in their disciplinary contexts: A conceptual analysis. *Studies in Higher Education, 27*(4), 405-417.

Newman, D. (2001). The academic achievement game: Designs of undergraduates' efforts to get grades. *Written Communication, 18*(4), 470-505.

North, S. (2003). Emergent disciplinarity in an interdisciplinary course: Theme use in undergraduate essays in the History of Science. Unpublished PhD. The Open University.

North, S. (2005a). Different values, different skills? A comparison of essay writing by students from arts and science backgrounds. *Studies in Higher Education, 30*, 517-533.

North, S. (2005b). Disciplinary variation in the use of Theme in undergraduate essays. *Applied Linguistics, 26*(3), 431-452.

O'Connell, F., & Jin, L. (2001). *A structural model of literature review: An analysis of Chinese postgraduate students' writing*. Paper presented at the BALEAP PIM conference: Focus on Chinese learners. Retrieved 01/11/2010, from http://www.baleap.org.uk/pimreports/2001/shu/abstracts.htm#secondary

Oakey, D. (2002). Formulaic language in English academic writing. In R. Reppen, S. Fitzmaurice & D. Biber (Eds.), *Using Corpora to Explore Linguistic Variation* (pp. 111-129). Amsterdam: John Benjamins.

Oakey, D. (2009). *An isotextual approach to comparisons of lexical bundles across disciplines*. Paper presented at the Aston Corpus Symposium. Retrieved 10/12/2010, from http://acorn.aston.ac.uk/Sym_Speakers09.html

Papp, S. (2009). Portsmouth Chinese-English Learner Corpus. Portsmouth: The University of Portsmouth.

Paquot, M. (2010). *Academic Vocabulary in Learner Writing: From Extraction to Analysis*. London: Continuum.

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Langage and Communication* (pp. 191-226). New York: Longman.

Pecorari, D. (2008). Repeated language in academic discourse: The case of Biology background statements. *Nordic Journal of English Studies, 7*(3), 9-33.

Pecorari, D. (2009). Formulaic language in Biology: A topic-specific investigation. In M. Charles, S. Hunston & D. Pecorari (Eds.), *At the Interface of Corpus and Discourse: Analysing Academic Discourses* (pp. 91-109). London: Continuum.

Petch-Tyson, S. (1998). Reader/writer visibility in EFL persuasive writing. In S. Granger (Ed.), *Learner English on Computer* (pp. 107-118). London/New York: Addison Wesley Longman.

Philip, G. (2008). Reassessing the canon: 'Fixed' phrases in general reference corpora. In S. Granger & F. Meunier (Eds.), *Phraseology* (pp. 95-108). Amsterdam: John Benjamins.

Pilcher, N., Cortazzi, M., & Jin, L. (2006). 'An immense variability'? British university supervisors' perceptions of mainland Chinese learners. *The East Asian Learner, 2*(2), 1-9.

Ping, D. N. F. (2007). Medium and learning in Chinese and English in Hong Kong classrooms. *Language Policy, 6*, 163-183.

Prior, P. (1998). *Writing/Disciplinarity: A Sociohistoric Account of Literate Activity in the Academy*. New Jersey: Lawrence Erlbaum.

Qi, L. (2004). *The Intended Washback of the National Matriculation English Test in China*. Guangzhou: Foreign Language Teaching and Research Press.

Qiufang, W., Maocheng, L., & Xiaoqin, Y. (2008). *Spoken and Written Corpus of Chinese Learners (SWECCL) 2.0*. Beijing: Foreign Language Teaching and Research Press.

Rai, L. (2008). Student writing in Social Work education. Unpublished PhD. The Open University.

Rayson, P. (2008a). From key words to key semantic domains. *International Journal of Corpus Linguistics, 13*(4), 519-549.

Rayson, P. (2008b). WMatrix. Lancaster: Computing Department, Lancaster University.

Renouf, A., & Sinclair, J. M. (1991). Collocational frameworks in English. In K. Aijmer & B. Altenberg (Eds.), *English Corpus Linguistics: Studies in Honour of Jan Svartvik* (pp. 128-143). London: Longman.

Richards, J. C., & Rodgers, T. S. (2001). *Approaches and Methods in Language Teaching* (2nd ed.). Cambridge/New York: Cambridge University Press.

Ringbom, H. (1998). Vocabulary frequencies in advanced learner English: A cross-linguistic approach. In S. Granger (Ed.), *Learner English on Computer* (pp. 41-52). London & New York: Addison Wesley Longman.

Römer, U. (2009). *The use of phraseological items in apprentice academic writing: Does nativeness matter?* Paper presented at the Aston Corpus Symposium.

Samraj, B. (2002). Introductions in research articles: Variations across disciplines. *English for Specific Purposes, 21*(1), 1-17.

Samraj, B. (2005). An exploration of a genre set: Research article abstracts and introductions in two disciplines. *English for Specific Purposes, 24*(2), 141-156.

Saw, S.-H., & Kesavapany, K. (2006). *Malaysia: Recent trends and challenges*. Singapore: Institute of Southeast Asian Studies.

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11*, 129-158.

Schmitt, N. (2004). *Formulaic Sequences: Acquisition, Processing and Use*. Amsterdam/Philadelphia: John Benjamins.

Schmitt, N., Grandage, S., & Adolphs, S. (2004). Are corpus-derived recurrent clusters psycholinguistically valid? In N. Schmitt (Ed.), *Formulaic Sequences: Acquisition, Processing and Use* (pp. 127-148). Amsterdam/Philadelphia: John Benjamins.

Schmitt, N., & McCarthy, M. (1997). *Vocabulary: Description, Acquisition and Pedagogy*. Cambridge: Cambridge University Press.

Scott, M. (2009). In search of a bad reference corpus. In D. Archer (Ed.), *What's in a Word-list? Investigating Word Frequency and Keyword Extraction* (pp. 79-92). Oxford: Ashgate.

Scott, M. (2010). WordSmith Tools. Liverpool: Lexical Analysis Software.

Scott, M., & Tribble, C. (2006). *Textual Patterns: Key Words and Corpus Analysis in Language Education*: John Benjamins.

Seidlhofer, B., & Widdowson, H. G. (2009). Accommodation and the idiom principle in English as a Lingua Franca. In K. Murata & J. Jenkins (Eds.), *Global Englishes in Asian Contexts* (pp. 26-39). Basingstoke: Palgrave Macmillan.

Sharpling, G. (2004). Inter-cultural issues in testing Chinese students' writing. *English Language Teacher Education and Development, 8*, 66-82.

Sharpling, G. (2010). When BAWE meets WELT: The use of a corpus of student writing to develop items for a proficiency test in grammar and English usage. *Journal of Writing Research, 2*(2), 179-195.

Shen, W. (2005). A study on Chinese student migration in the United Kingdom. *Asia Europe Journal, 3*, 429-436.

Shu, L. (2006). The successors to Confucianism or a new generation? A questionnaire study on Chinese students' culture of learning English. *Language, Culture & Curriculum, 19*(1), 122-147.

Simpson, R. C. (2004). Stylistic features of academic speech: The role of formulaic expressions. In U. Connor & T. A. Upton (Eds.), *Discourse in the Professions: Perspectives from Corpus Linguistics* (pp. 37-64). Amsterdam/Philadelphia: John Benjamins.

Sinclair, J. M. (1991). *Corpus Concordance Collocation*. Oxford: Oxford University Press.

Singapore Ministry of Education. (2011). *Returning Singaporeans: Mother-Tongue Language Policy*. Retrieved 04/04/2011, from http://www.moe.gov.sg/education/admissions/returning-singaporeans/mother-tongue-policy/.

Sorapure, M. (2010). Information visualization, Web 2.0, and the teaching of writing. *Computers and Composition, 27*, 59-70.

Speelman, D., Tummers, J., & Geeraerts, D. (2009). Lexical patterning in a construction grammar: The effect of lexical co-occurrence patterns on the inflectional variation in Dutch attributive adjectives. *Constructions & Frames, 1*(1), 87-118.

Sperberg-McQueen, C. M., & Burnard, L. (Eds.) (2004). *Guidelines for Electronic Text Encoding and Interchange: TEI P3*. Chicago and Oxford: The Text Encoding Initiative. (Reprinted 1999.) Retrieved 10/08/09, from http://www.tei-c.org/Guidelines/index.htm.

Spöttl, C., & McCarthy, M. (2004). Comparing knowledge of formulaic sequences across L1, L2, L3, and L4. In N. Schmitt (Ed.), *Formulaic Sequences: Acquisition, Processing and Use* (pp. 191-219). Amsterdam/Philadelphia: John Benjamins.

Stapleton, P. (2010). Writing in an electronic age: A case study of L2 composing processes. *Journal of English for Academic Purposes, 9*(4), 295-307.

Street, B. (1998). New literacies in theory and practice: What are the implications for language in education? *Linguistics and Education, 10*(1), 1-24.

Stubbs, M., & Barth, I. (2003). Using recurrent phrases as text-type discriminators. *Functions of Language, 10*(1), 61-104.

Swales, J. (1990). Nonnative speaker graduate engineering students and their introductions: Global coherence and local management. In U. Connor & A. M. Johns (Eds.), *Coherence in Writing: Research and Pedagogical Perspectives* (pp. 189-207). Alexandria VA: TESOL.

Swan, M., & Smith, B. (2001). *Learner English: A Teacher's Guide to Interference and Other Problems*. Cambridge: Cambridge University Press.

Tallack, D. (2006). *Internationalisation: The University of Nottingham's campus in China*. Paper presented at the 2nd Biennial International Conference, University of Portsmouth. Retrieved 14/05/11 from www.lass.soton.ac.uk/education/CLearnConfRpt.doc

Tang, R. (2009). A dialogic account of authority in academic writing. In M. Charles, S. Hunston & D. Pecorari (Eds.), *At the Interface of Corpus and Discourse: Analysing Academic Discourses* (pp. 170-188). London: Continuum.

Tang, R., & John, S. (1999). The 'I' in identity: Exploring writer identity in student academic writing through the first person pronoun. *English for Specific Purposes, 18* (Supplement), S23-S39.

Thewissen, J. Y. (forthcoming). Accuracy across L1 backgrounds and proficiency levels: Insights from an error-tagged EFL learner corpus. Unpublished PhD. Université Catholique de Louvain.

Thewissen, J. Y., Bestgen, J., & Granger, S. (2006). *Using error-tagged learner corpora to create English-specific CEF descriptors*. Paper presented at the Third Annual Conference of EALTA.

Thompson, P. (2001). A pedagogically-motivated corpus-based examination of PhD theses: Macrostructure, citation practices and uses of modal verbs. Unpublished PhD. University of Reading.

Thompson, P. (2005). Points of focus and position: Intertextual reference in PhD theses. *Journal of English for Academic Purposes, 4*, 307-323.

Thompson, P. (2007). Editorial. *Journal of English for Academic Purposes, 6*(4), 285-288.

Thompson, P. (2009). Shared disciplinary norms and individual traits in the writing of British undergraduates. In M. Gotti (Ed.), *Commonality and Individuality in Academic Discourse* (pp. 53-82). Bern: Peter Lang.

Tian, J. (2008). The influence of undergraduate learning contexts on Chinese graduate students' argumentation and critical thinking in writing. Unpublished PhD. University of York.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*: John Benjamins.

Tomlinson, B. (2005). The future for ELT materials in China. *Electronic Journal of Foreign Language Teaching, 2*(2), 5-13. Retrieved 04/04/2011, from http://e-flt.nus.edu.sg/v2n22005/tomlinson.pdf

Tono, Y. (2009). Learner corpus analysis and SLA. In P. Baker (Ed.), *Contemporary Corpus Linguistics* (pp. 184-203). London/New York: Continuum.

Tremblay, A. (2009). Processing advantages of lexical bundles: Evidence from selfpaced reading, word and sentence recall, and free recall with event-related brain potential recordings. Unpublished PhD. University of Alberta.

Tribble, C. (2009). Writing academic English: A survey review of current published resources. *English Language Teaching Journal, 63*(4), 400-417.

Tucker, G. (2005). Extending the lexicogrammar: Towards a more comprehensive account of extraclausal, partially clausal and non-clausal expressions in spoken discourse. *Language Sciences, 27*(6), 679-709.

UK Council for International Student Affairs (UKCISA). (2011). Higher Education Statistics. Retrieved 11/2/2011, from http://www.ukcisa.org.uk/about/statistics_he.php

Underwood, G., Schmitt, N., & Galpin, A. (2004). The eyes have it: An eye-movement study into the processing of formulaic sequences. In N. Schmitt (Ed.), *Formulaic Sequences: Acquisition, Processing and Use* (pp. 153-168). Amsterdam/Philadelphia: John Benjamins.

Van Leeuwen, T. (2005). *Introducing Social Semiotics*. London: Routledge.

Wang, W., & Wen, Q. (2002). L1 use in the L2 composing process: An exploratory study of 16 Chinese EFL writers. *Journal of Second Language Writing, 11*(3), 225-246.

Wang, X. (2003). *Education in China Since 1976.* Jefferson, N. Carolina: McFarland & Company.

Warburton, N. (2006). *The Basics of Essay Writing*. London: Routledge.

Wen, Q. F., & Ding, Y. R. (2004). A study of frequency adverbs used by advanced English learners in China. *Modern foreign languages, 2*, 141-147.

Wen, Q. F., Ding, Y. R., & Wang, W. Y. (2003). Features of oral style in English compositions of advanced Chinese EFL learners. *Foreign Language Teaching and Research, 4*, 268-274.

Wen, W. P., & Clement, R. (2003). A Chinese conceptualisation of willingness to communicate in ESL. *Language, Culture & Curriculum, 16*(1), 18-38.

Whitley, E. (2007). *Student academic writing in the internet age: Studying diversity in practice*. Paper presented at the 28th International Conference on Information Systems.

Wiechmann, D., & Fuhs, S. (2006). Concordancing software. *Corpus Linguistics and Lingustic Theory, 2*(1), 107-127.

Wiktorsson, M. (2000). The production rate of *prefabs*: A pilot study. In M. Aparici (Ed.), *Working papers in developing literacy across genres, modalities and languages* (Vol. 3, pp. 225-234). Barcelona: University of Barcelona.

Wiktorsson, M. (2003). *Learning idiomaticity: A corpus-based study of idiomatic expressions in learners' written production*. Lund: Lund University.

Wood, D. (2004). An empirical investigation into the facilitating role of automatized lexical phrases in second language fluency development. *Journal of Language and Learning, 2*(1), 27-50.

Wood, D. (2007). Mandarin Chinese speakers in a study abroad context: Does acquisition of formulaic sequences facilitate fluent speech in English? *The East Asian Learner, 3*(2).

Woodward-Kron, R., & Jamieson, H. (2007). Tensions in the writing support consultations: Negotiating meanings in unfamiliar territory. In C. Gitsakis (Ed.), *Language and Languages: Global and Local Tensions* (pp. 40-61). Newcastle: Cambridge Scholars Publication.

Wray, A. (2002). *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

Wray, A. (2008). *Formulaic Language: Pushing the Boundaries*. Oxford: Oxford University Press.

Wray, A., & Namba, K. (2003). Use of formulaic language by a Japanese-English bilingual child: A practical approach to data analysis. *Japan Journal for Multilingualism and Multiculturalism, 9*(1), 24-51.

Wray, A., & Perkins, M. R. (2000). The functions of formulaic language: An integrated model. *Language & Communication, 20*, 1-28.

Wyatt-Smith, C., & Kimber, K. (2009). Working multimodally: Challenges for assessment. *English Teaching: Practice and Critique, 8*(3), 70-90.

Xiao, R., & McEnery, T. (2005). Two approaches to genre analysis: Three genres in modern American English. *Journal of English Linguistics, 33*(1), 62-82.

Yang, H. (2005). Corpus-based Analysis of Chinese Learner English. *Journal of Foreign Languages in China, 2*(6), 1-20.

Yeung, L. (2007). In search of commonalities: Some linguistic and rhetorical features of business reports as a genre. *English for Specific Purposes, 26*(2), 156-179.

Zhang, G. X., Li, L. M., & Suen, L. (2005). *Chinese in Steps*. London: Cypress Books.

Zhu, W. (2004). Writing in business courses: An analysis of assignment types, their characteristics, and required skills. *English for Specific Purposes, 23*(2), 111-135.

Appendix A ICLE and BAWE titles

ICLE titles

The following are the suggested essay titles from the ICLE site (http:// www.uclouvain.be/en-317608.html):

1. Crime does not pay.
2. The prison system is outdated. No civilized society should punish its criminals: it should rehabilitate them.
3. Most university degrees are theoretical and do not prepare students for the real world. They are therefore of very little value.
4. A man/woman's financial reward should be commensurate with their contribution to the society they live in.
5. The role of censorship in Western society.
6. Marx once said that religion was the opium of the masses. If he was alive at the end of the 20th century, he would replace religion with television.
7. All armies should consist entirely of professional soldiers: there is no value in a system of military service.
8. The Gulf War has shown us that it is still a great thing to fight for one's country.
9. Feminists have done more harm to the cause of women than good.
10. In his novel Animal Farm, George Orwell wrote ‘All men are equal: but some are more equal than others’. How true is this today?
11. In the words of the old song ‘Money is the root of all evil’.
12. Europe.
13. In the 19th century, Victor Hugo said: ‘How sad it is to think that nature is calling out but humanity refuses to pay heed. ‘Do you think it is still true nowadays?
14. Some people say that in our modern world, dominated by science technology and industrialization, there is no longer a place for dreaming and imagination. What is your opinion?

BAWE titles

The following are a sample of BAWE titles from different disciplines, and year groups of study:

1. Using your hypothetical case study EITHER explain the relevance of different motivational theories to OR explore the role of communication in the situation described. (year 1, Agriculture).
2. Experiment 5- ELISA (year 1, Biology).
3. A-Z cloth management report on the costs and quantities of yarn produced (year 1, Business).
4. Consider how Victorian notions of women's madness affected the lives of women in class specific ways (year 1, Sociology).
5. A study of population growth within the predator prey model (year 2, Mathematics).
6. What is the Phillips curve? Explain why critics believe the ratio no longer holds. (year 2, Economics).
7. A report into the current and future financial and market position of the Markstrat company *Jolly SU*. (year 2, Business).
8. 'If a realistic medical jurisprudence is to develop, judges must extend their focus beyond rights and duties and confront the fundamental issue of resources.' Discuss (year 2, Law).
9. Legal issues surrounding open source and open standards (year 3, Computer Science).
10. DSP in digital communications (year 3, Cybernetics and Engineering).
11. Exercise 11 Microbiological risks associated with mail delivery of vacuum packed smoked salmon (year 3, Food Sciences).
12. Reflections of the likelihood of my becoming an entrepreneur (year 3, HLTM).

Appendix B Anonymized example of BAWE form

The Corpus of British Academic Written English (BAWE)

Surname: ..Xxxxxxx. Forename: ....Xxxxxx 3135a

Male/Female Date of birth: *23/01/86* Email address: *xxxx@brookes.ac.uk*

Your first language: *English*.....

Your secondary education (since 11 years old but before university) was:

X All in UK	All overseas.	Some in UK, some overseas. Please state number of years in UK

Your year of study (when you wrote the assignment):

first year UG X	second year UG	third year UG no intercalated year	third year UG with intercalated year	fourth year UG	PG (masters level or diploma)

Other (please specify): .....

Your home department: ..... *Anthropology*

Course of study: *BA Anthropology and Sociology*

Brief Title of assignment: *Critical Review of 'Our Land Was A Forest' ('the Assignment')*

Year & Month when assignment written: ....*Jan 05*

Module title: ... *Minorities and Marginalities: Class and Conflict in Japan*

Module tutor's name: .... *Xxxxxx Xxxxxx*

Module code: .. *U20137*

Grade/Mark received: 65%

Was this work co-authored (e.g. group assignment)? .....no.....

Please indicate the type of assignment, according to your understanding of the task, by choosing 1 of the options below:
Case-Study/Essay/Exercise/Notes/Presentation/Report/Review/none of the above (please specify):

In consideration of the sum of £3.00 paid by Oxford Brookes University ('the University') to me the receipt of which I hereby acknowledge I hereby *assign* to the University the Assignment and my intellectual property rights in the same together with waiving my moral rights. I warrant that the Assignment is my own work and agree that I shall advise the

University of the quotation or inclusion in the Assignment of any textual or illustrative material taken from other sources and provide the University with full details of the original source of that material.

I acknowledge that the Assignment may be submitted to the JISC Plagiarism Advisory Service (a facility which carries out electronic comparison of students' work against electronic sources, including work submitted by students at other institutions).

..... (Signature) ..... (Date)

Appendix C Genre families in the BAWE corpus

Adapted from Heuboeck et al., 2008

Genre Families	Social purpose/ Components/ Genre network	Genres (examples from each family)
Case Study	- To gain an understanding of professional practice through the analysis of a single exemplar - Description of a particular case, often multifaceted, with recommendations or suggestions for future action. - Typically corresponds to professional genres (e.g. in Business, Engineering)	Business start-up Company report (starts with executive summary) Investigation report Organisation analysis Single issue Tourism report
Critique	- To demonstrate understanding of the object of study; to demonstrate the ability to evaluate and/or assess the significance of the object of study - Includes descriptive account, explanation, and evaluation; often involves tests. - May correspond to part of a research paper, professional design specification or expert evaluation.	Academic paper review Business/organisation analysis Business/organisation evaluation Interpretation of results Product/building evaluation Project evaluation System evaluation
Design specification	- To demonstrate the ability to design a product or procedure that could be manufactured or implemented. - Typically includes design brief, design considerations, and design plan; may include development and testing of design. - May correspond to a professional design specification, or to part of a proposal or research report.	Application design Product design System design Database design
Empathy writing	- To demonstrate understanding and appreciation of the relevance of academic ideas by translating them into a non-academic register, to communicate to a non-specialist readership - May be formatted as a letter, newspaper article or similar non-academic genre	Expert advice to industry Expert advice to lay person Information leaflet Job application Letter (e.g. reflective letter to friend; business letter,

	- May correspond to professional writing.	newspaper article)
Essay	- To develop the ability to construct a coherent argument and develop critical thinking skills. - Discussion, exposition, factorial, consequential, challenge, or commentary. - May correspond to a published academic/specialist paper.	Challenge Commentary Comparison Consequential Discussion Exposition Factorial
Exercise	- To provide practice in key skills (e.g. the ability to interrogate a database, perform complex calculations, or explain technical terms or procedures), and to consolidate knowledge of key concepts. - Data analysis or a series of responses to questions. - May correspond to part of report or research paper.	Calculations Data analysis Mixed (e.g. calculations + essays) Short answers Statistics exercise
Explanation	- To demonstrate understanding of the object of study; and the ability to describe and/or assess its significance. - Includes descriptive account, explanation. - May correspond to a published explanation, or to part of a research paper or professional design specification.	Business review Methodology review Product development overview Species/breed overview Substance/phenomenon overview System/process overview
Literature survey	- To demonstrate familiarity with literature relevant to the focus of study. -Includes summary of literature relevant to the focus of study and varying degrees of critical evaluation. - May correspond to a published paper or anthology, or to part of a research paper.	Annotated bibliography Literature review Notes taken from multiple sources Summary book chapter
Methodology recount	- To become familiar with disciplinary procedures and methods, and additionally to record experimental findings. - Describes procedures undertaken by writer; may include Introduction, Methods, Results and Discussion sections, or these functions may be	Computer analyses Data analysis report Lab report Materials selection report Development report

	realized iteratively. - May correspond to a section within a research report or research paper.	
Narrative recount	- To develop awareness of motives and/or behaviour in individuals (including self) or organisations - Fictional or factual recount of events - May correspond to published literature, a professional proposal or a report, or to part of a research paper.	Account of literature search Account of website search Character outline Reflective recount
Problem question	- To practise applying specific methods in response to simulated professional problems - Problem (may not be stated in assignment); application of relevant arguments or presentation of possible solution(s) in response to scenario - Problems or situations may resemble or be based on real legal, engineering, accounting or other professional cases	Logistics simulation Medical problem
Proposal	- To demonstrate ability to make a case for future action - Includes purpose, detailed plan, persuasive argumentation - May correspond to professional or academic proposals	Business plan Design proposal Marketing plan Research proposal Catering plan
Research report	- To demonstrate ability to undertake a complete piece of research including research design, and an appreciation of its significance in the field - May include Literature Review, Methods, Findings, Discussion; or may include several ‘chapters’ relating to the same theme - May correspond to a published experimental research paper or topic-based research paper	Research paper Topic-based dissertation

Appendix D: Keywords in Chi123

All keywords are given in descending order of keyness (RC = Eng123).

Key:	*keyword*	connectors e.g. *in the long run*
	KEYWORD	informal items e.g. *besides*
	KEYWORD	keyword containing a pronoun e.g. *we need*
	keyword	references to data or visuals *e.g. according to*

Keywords and key 2-grams in Chi123 occur at least 20 times (71.5pmw) in the texts of at least five individuals, in both years 1/2 and 3, and within at least three disciplines.

Keywords

corporate, curve, rate, responsibility, output, ***WE****, price, samples, population, BESIDES, capital, real, #, students, guidelines, media, economy, level, growth, firms, might, income, fault, while, film, noise,* ***formula***[37]*, model,* ***according****, channel, restriction, competitors, portfolio, variance, is, programming, higher, generation, among, nowadays, same, mode, demand, viability, care, competitive, exchange, bits, industry, nuclear, error, different, languages, supply,* ***figure****, value, behaviour, liability, electrical, strategic, foreign, mm, liquid, bit, connect, signal, people, meanwhile, standard, strategies, internet, managers, wage, digested, mass, term, products, relationship, directors, sequence, performance, temperature, countries, words, parameters, density, volume, shift, note, clothing, content, got, consumption, slope, recognized, customers, motor, long, experiment, market, culture, fixed, blocks, M, measurement, Chinese, collapse, after, random, nominal, statistical, reaction, search, depends, product, about, organizations, mechanical, diminishing, new, LOTS, measured, deviation, in, marginal, investment, bottom, services, on, find, high, firm, interviews, capabilities, storage, free, sample, nevertheless,* ***eq****, digital, cultures, kind, time, decrease,* ***refer****, moisture, gas, margin, normal, approaches, detect, currency, program, […] what's* (19 occurrences)

Key 2-grams

of care, non state, real time, of population, ***according to****, the mass,* ***as below****, find out, people can,* ***is formula****, long term, this model, might be, which is, among the, the real, other words, all the, is about, is the, curve of, the same, long run, this experiment, out the, and services, the output, in this, the response, economic growth,* ***WE WILL****, of capital, available at, curve is, kind of,* ***the appendix****, other hand, the long, the slope, on the, relationship between, depends on, different from,* ***the figure****, so on, while the, is called, as well, becomes a, in addition, higher than, growth rate, level of,* ***refer to, standard deviation****, model is, in real, goods and, choice of, same as, the fixed,* ***referring to****, on is, LOTS OF, based on, rate of, of them, by using, the whole, the relationship,* ***the equation****, with different, important role, value of, can not, and output,* ***illustrated in****,* ***WE CAN****, output and, which means, of business,* ***WE COULD****, in other, change the, the result, United States, the reaction, the economy, short term, the industry,* ***formula and****, is set,* ***WE NEED****, when there,* ***formula is****, the total, than the, change of, a firm, same time, the mean, performance*

37 *Formula* throughout this appendix has been given in lower case to distinguish it from key categories indicated through capitals. In the rest of the thesis, upper case is used to denote this item as it replaces a range of formulae in BAWE tagging.

and, the customers, to detect

[…] what's more (13 occurrences)

Key 3-grams and 4-grams occurring six or more times (21.4pmw)

Key 3-grams	Freq.	%	RC Freq.	RC %	Keyness
according to the	73	0.03	77		77
to find out	35	0.01	18		62
in this experiment	44	0.02	37		57
in other words	29	0.01	13		55
and so on	23		9		46
the change of	13		0		46
the other hand	56	0.02	81		42
refer to the	29	0.01	22		40
but not least	11		0		39
a kind of	17		5		38
on the other	65	0.02	112		38
based on the	71	0.03	134	0.01	35
last but not	10		0		35
there might be	10		0		35
in the long	32	0.01	37		31
this kind of	16		7		31
A LITTLE BIT	8		0		28
higher than the	31	0.01	39		27
goods and services	18		12		27
the relationship between	45	0.02	77		26
the long run	22		23		24
when there is	18		15		23
the long term	32	0.01	48		23
the slope of	13		7		22
is the same	23		27		22
at that time	14		9		22
products and services	14		9		22
depends on the	33	0.01	53		21
is higher than	15		11		21
in the appendix	22		26		21
important role in	19		20		20

Key 4-grams	Freq.	%	RC Freq.	RC %	Keyness
on the other hand	54	0.02	81		38
LAST BUT NOT LEAST	10		0		35
is a kind of	8		0		28
according to the equation	7		0		25
in the long run	19		19		21

WE ALSO NEED TO	6		0	21
is the same as	13		9	19
in other words the	9		3	19
in the end of	8		2	19
at the same time	35	0.01	65	18
all over the world	9		4	17
in order to avoid	14		13	17

Key 5-gram	**Freq.**	**RC. Freq.**	**Keyness**
is the same as the	7	0	25

All negative keywords and n-grams (using the same search parameters)

Negative Keywords

*that, this, would, be, been, criminal, women, whilst, these, have, within, being, land, upon, class, must, a, organic, seen, review, justice, to, disease, an, project, not, was, environmental, had, therefore, **HIS**, allow, political, material, could, stress, species, where, health, made, however, evidence, may, society, areas, costs, research, due, act, flow **THEIR**, **THEY**, page, as*

Negative Key 2-grams

*this is, within the, have been, to be, of the, the project, would have, it was, that the, would be, this was, that it, as it, that of, be seen, for the, that this, **THAT THEY**, fact that, means that, also be, due to, must be, has been, et al, on a, able to, likely to, and that, **AS THEY**, that are, it would, the business, right to, for the*

Negative Key 3-grams	**Freq.**	**%**	**RC Freq.**	**RC &**	**Keyness**
can be seen	30	0.01	389	0.03	-37
the fact that	26		298	0.02	-23
this is a	9		162	0.01	-23
this means that	11		169	0.01	-20
in the UK	25		272	0.02	-20

Negative Key 4-grams	**Freq.**	**%**	**RC Freq.**	**RC %**	**Keyness**
can be seen in	7		133		-20
can be seen that	5		95		-14
the fact that the	5		91		-13
to be able to	5		84		-11

(No negative key 5-grams were found)

Appendix E Normalized and raw counts for Chapter 6

Informal items

	Chi12		Chi3		Eng12		Eng3	
	pmw	raw	pmw	raw	pmw	raw	pmw	raw
lots/a lot	271	(38)	208	(29)	145***	(127)	76	(35)
besides	235*	(33)	122	(17)	8	(7)	4	(2)
a (little)bit	71	(10)	72	(10)	10	(17)	9	(4)
last but not least	50	(7)	22	(3)	0	(0)	0	(0)
contracted verb forms	143	(20)	115	(16)	231	(203)	190	(87)

Connectors

	Chi12		Chi3		Eng12		Eng3	
	pmw	raw	pmw	raw	pmw	raw	pmw	raw
on the other hand	257*	36	129	18	59	52	63	29
besides	228*	32	122	17	8	7	4	2
at the same time	185**	26	65	9	47	41	52	24
nevertheless	157	22	179	25	62	54	41	19
nowadays	150	21	86	12	2	2	15**	7
in the long run	114**	16	22	3	13	11	17	8
and so on	114	16	50	7	5	4	11	5
in other words	100	14	108	15	7	6	15	7
at that time	86**	12	14	2	7	6	35	16
meanwhile	86	12	72	10	2	2	9	4
what's more	64	9	29	4	0	0	0	0
last but not least	50	7	22	3	0	0	0	0
however	1290	181	1536	214	1887	1655	2071*	950
therefore	1140*	160	868	121	1504	1319	1543	708

Number of visuals and lists by year group (pmw)

pmw	Tables	Figures	Formulae	Lists	Listlikes
Chi12	720	976	2929	456	3263
Chi3	782	1458***	3093	344	4248****
Eng12	472**	780	1942*	425***	1424
Eng3	353	817	1742	294	1364

Significance figures

All comparisons are between year group corpora i.e. Chi12 and Chi3; Eng12 and Eng3.

* 95th percentile; 5% level; p<.05; critical value = 3.84
** 99th percentile; 1% level; p<.01; critical value = 6.63
*** 99.9th percentile; 0.1% level; p<.001; critical value = 10.83
**** 99.99th percentile; 0.01% level; p<.0001; critical value = 15.13

Appendix F Classification of 4-grams

Key to functional classification (based on Hyland, 2008a,b)

Function	Specific function	Code
Participant-oriented	**Engagement features**, address readers directly,	p-e
	Stance features, convey the writer's attitudes and evaluations	p-s
Research-oriented	**Description**	r-d
	Location, indicating time/place	r-l
	Procedure	r-p
	Quantification	r-q
	Topic, related to the field of research	r-t
Text-oriented	**Framing signals**, situate arguments by specifying limiting conditions	t-f
	Resultative signals, mark inferential or causative relations between elements	t-r
	Structuring signals, text-reflexive markers which organize stretches of discourse or direct reader elsewhere in text	t-s
	Transition signals, establishing additive or contrastive links between elements	t-t

Key to structural classification (based on Biber et.al., 1999)

Broad category	Structural pattern
NP-based	(1) NP + *of*-phrase fragment
	(2) NP + other post-modifier fragment
	(3) pronoun/NP (+aux) + *be*
PP-based	(4) PP + embedded *of*-phrase fragment
	(5) other PP fragment
VP-based	(6) anticipatory *it* + VP/AdjP (+complement clause)
	(7) passive verb + PP fragment
	(8) *be* + NP/AdjP
	(9) (NP+) (verb +) *that*-clause fragment
	(10) (V/Adj+) *to*-clause fragment
Other	(11) other expressions

CHI 12	str.	func.	no.		CHI 3	str.	func.	no.
ON THE OTHER HAND	5	t-t	36		TO FIND OUT THE	10	r-p	20
AT THE SAME TIME	5	r-l	26		ON THE OTHER HAND	5	t-t	18
THE RELATIONSHIP BETWEEN	2	r-d	17		AS A RESULT OF	4	t-r	14
AS A RESULT OF	4	t-r	17		AS WELL AS THE	11	t-t	13
IN THE CASE OF	1	t-f	16		CAN BE USED TO	7	r-p	12
IN THE LONG RUN	5	t-t	16		IN THE CASE OF	4	t-f	11
PLAY(S)/PLAYED AN IMPORTAN	8	r-d	12		IT IS IMPORTANT TO	6	p-e	9
IS ONE OF THE (MOST)	8	r-q	11		AT THE SAME TIME	5	r-l	9
AS WELL AS THE	11	t-t	11		AT THE END OF	4	r-l	9
THE SENSITIVITY OF THE	1	r-d	10		IT CAN BE SEEN THAT)	6	p-e	8
IN ORDER TO ACHIEVE	5	r-p	10		IT IS POSSIBLE TO	6	p-e	8
IN TERMS OF THE	4	t-f	10		IN THE FORM OF	4	r-d	8
(IS DUE) TO THE FACT THAT	8	p-s	9		IT IS NECESSARY TO	6	p-e	7
IT IS BELIEVED THAT	6	p-s	9		IT IS EASY TO	6	p-e	7
WILL BE ABLE TO	10	p-s	8		IS THE SAME AS	8	r-d	7
IS BASED ON THE	7	p-s	8		FOR EACH OF THE	5	r-q	7
IS PROPORTIONAL TO THE	8	r-d	8		IN ORDER TO BE	5	t-r	7
IS DETERMINED BY THE	7	r-d	8		WILL BE ABLE TO	10	p-s	6
IN ORDER TO AVOID	5	r-p	8		THE QUALITY OF THE	1	r-d	6
AS ONE OF THE	4	t-f	8		AT THE BEGINNING OF	4	r-l	6
TO THE FACT THAT	5	p-s	7		IN THE END OF	4	r-l	6
THAT THERE IS A	9	r-d	7		IN ORDER TO AVOID	5	r-p	6
IS A KIND OF	8	r-d	7		IS ONE OF THE	8	r-q	6
THE NATURE OF THE	1	r-d	7		A WIDE RANGE OF	1	r-q	6
THE END OF THE	1	r-l	7		IN TERMS OF THE	4	t-f	6
WAS ONE OF THE (MOST)	8	r-q	7		AS SHOWN IN FIGURE	5	t-s	6
A LARGE NUMBER OF	1	r-q	7		IN THE SAME WAY	5	t-t	6
ALL OVER THE WORLD	5	r-t	7		CAN BE SEEN THAT	9	p-e	5
LAST BUT NOT LEAST	11	t-t	7		IT IS CLEAR THAT	6	p-s	5
IT IS IMPORTANT TO	6	p-e	6		THE PERFORMANCE OF THE	1	r-d	5
TO DEAL WITH THE	10	p-s	6		THE SLOPE OF THE	1	r-d	5
IS THE SAME AS	8	r-d	6		AN IMPORTANT ROLE IN	2	r-d	5
THE SIZE OF THE	1	r-d	6		THE SAME AS THE	2	r-d	5
TO THE DEVELOPMENT OF	1	r-d	6		THE END OF THE	1	r-l	5
AT THE END OF	4	r-l	6		IN ORDER TO FIND	5	r-p	5
USED IN THIS EXPERIMENT	10	r-p	6		OF THE NUMBER OF	1	r-q	5
CAN BE USED TO	7	r-p	6		THAT THERE IS NO	9	r-q	5
(THERE) ARE TWO TYPES OF	8	r-q	6		THE LIFE OF THE	1	r-t	5
THE AIM OF THE EXPERIMENT(S	1	t-f	6		AS LONG AS THE	5	t-f	5
IT IS DIFFICULT TO	6	p-e	5		IN THE SHORT TERM	5	t-t	5
IT IS ESSENTIAL TO	6	p-e	5		IS CONSISTENT WITH THE	8	p-s	4
IT IS VERY IMPORTANT TO	6	p-e	5		CAN BE USED AS	7	p-s	4
WE CAN SEE THAT	3	p-e	5		IS EQUAL TO THE	8	r-d	4
IS ALSO A	8	p-s	5		THE SIZE OF THE	1	r-d	4
THE SURFACE OF THE	1	r-d	5		THE BEGINNING OF THE	1	r-l	4
IN THIS CASE THE	5	r-p	5		IN ORDER TO ACHIEVE	5	r-p	4
THERE ARE A NUMBER OF	3	r-q	5		IN ORDER TO MAINTAIN	5	r-p	4
THE MAJORITY OF THE	1	r-q	5		IN THIS CASE THE	5	r-p	4
OF GOODS AND SERVICES	5	r-t	5		ONE OF THE MOST	1	r-q	4
WITH RESPECT TO THE	5	t-f	5		IN THE NEXT SECTION	5	t-s	4

ENG 12	str.	func.	no.		ENG 3	str.	func.	no.
IN THE CASE OF	4	t-f	101		CAN BE SEEN IN	7	t-s	77
IN THE FORM OF	4	r-d	83		AS A RESULT OF (THE)	4	t-r	58
THE END OF THE	1	r-l	77		IT IS POSSIBLE TO	6	p-e	53
AS A RESULT OF	4	t-r	77		IT CAN BE SEEN (THAT)	6	p-e	42
CAN BE USED TO	7	r-p	76		IN THE CASE OF	4	t-f	41
IT CAN BE SEEN THAT	6	p-e	72		CAN BE USED TO	7	r-p	38
CAN BE SEEN THAT	9	p-e	61		IT IS IMPORTANT TO	6	p-e	37
THE FACT THAT THE	2	t-t	60		AS WELL AS THE	11	t-t	32
TO BE ABLE TO	10	r-p	59		IN THE FORM OF	4	r-d	31
CAN BE SEEN IN	7	t-s	56		AT THE END OF	4	r-l	31
IT IS IMPORTANT TO	6	p-e	54		THE FACT THAT THE	2	t-t	31
ON THE OTHER HAND	5	t-t	52		ONE OF THE MOST	1	r-q	30
AT THE END OF	4	r-l	51		TO THE FACT THAT	11	t-t	30
IT IS CLEAR THAT	6	p-s	50		ON THE OTHER HAND	5	t-t	29
IT IS POSSIBLE TO	6	p-e	49		THE WAY IN WHICH (THE)	2	t-f	28
THIS IS DUE TO	3	p-s	44		THE END OF THE	1	r-l	27
THIS MEANS THAT THE	9	p-s	44		DUE TO THE FACT	11	p-s	25
THAT THERE IS A	9	r-d	42		TO BE ABLE TO	10	r-p	25
AS WELL AS THE	11	t-t	42		THE DEVELOPMENT OF THE	1	r-d	24
AT THE SAME TIME	5	r-l	41		AT THE SAME TIME	5	r-l	24
CAN BE SEEN AS	7	p-e	39		TO ENSURE THAT THE	10	r-p	24
IT IS DIFFICULT TO	6	p-s	39		THE EXTENT TO WHICH	2	t-f	24
THE SIZE OF THE	1	r-d	38		ARE MORE LIKELY TO	10	p-s	23
THE REST OF THE	1	r-q	38		THE ROLE OF THE	1	r-p	23
ONE OF THE MOST	1	r-q	37		THE RESULTS OF THE	1	t-f	23
WILL NEED TO BE	10	r-p	36		IT IS CLEAR THAT	6	p-s	22
THIS IS BECAUSE THE	3	p-s	35		IS ONE OF THE (MOST)	8	r-q	21
WILL BE ABLE TO	10	p-s	34		IN TERMS OF THE	4	t-f	21
THE TOP OF THE	1	r-d	34		THE REST OF THE	1	r-d	20
IN THE ABSENCE OF	4	t-f	34		IN THE DEVELOPMENT OF	4	p-s	19
THE EXTENT TO WHICH	2	t-f	34		THIS IS DUE TO (THE)	3	p-s	19
TO THE FACT THAT	11	p-s	33		THE SURFACE OF THE	1	r-d	19
IS LIKELY TO BE	10	p-s	32		IT IS NECESSARY TO	6	p-e	18
WOULD HAVE TO BE	10	p-s	32		IS DUE TO THE	8	p-s	18
CAN BE FOUND IN	7	t-s	32		BE SEEN AS A	7	p-e	17
THE NATURE OF THE	1	r-d	31		THE NATURE OF THE	1	r-d	17
IS ONE OF THE	8	r-q	31		THE PERFORMANCE OF TH	1	r-d	17
IN TERMS OF THE	4	t-f	31		THE CENTRE OF THE	1	r-d	16
NOT BE ABLE TO	10	p-s	30		THE USE OF THE	1	r-d	16
THE RELATIONSHIP BETWEE	2	r-d	28		THERE HAS BEEN A	3	r-d	16
TO ENSURE THAT THE	10	r-p	28		THE INTRODUCTION OF TH	1	t-t	16
IN THIS CASE THE	5	t-f	28		THAT THERE IS A	9	p-s	15
THE CASE OF THE	1	t-f	28		THE SIZE OF THE	1	r-d	15
AND AS A RESULT	5	t-r	28		ARE LIKELY TO BE	10	p-s	14
A WIDE RANGE OF	1	r-q	27		THE SUCCESS OF THE	1	r-d	13
THE MAJORITY OF THE	1	r-q	26		A MEMBER OF THE	1	r-d	13
THE WAY IN WHICH	2	t-f	26		IN RELATION TO THE	5	r-d	13
(DUE) TO THE FACT (THAT)	11	p-s	24		THROUGH THE USE OF	4	r-d	13
ON THE BASIS OF	4	t-r	24		FOR THE PURPOSE OF	4	r-p	13
THAT THERE WAS A	9	r-d	23		IN THE ABSENCE OF	4	t-f	13

Appendix G Keywords in three disciplines

Biology

Key in Chi-Biology	*primers, PCR, methylation, dimethoate, population, restriction, alignment, real, digestion, rate, sequences, eggshell, samples, might, phylogeny, salinity, template, tree, enzymes, spartum, level, simulation, tangere, noli, dispersion, very, madidus, genomic, T*
Key in Eng-Biology	*of, in, that, are, by, from, which, cell, protein, cells, these, they, been, may, were, when, two, water, between, found, species, plants, gene, al, due, et, both, different, C, form, figure, site, genes, light, present, enzyme, growth, DNA, plant, B, acid, during, shown, activity, concentration, specific, rate, similar, PH, par, energy, amount, produced, experiment, reaction, study, L, pie, size, addition, SKN, soil, carried, sequence, absorbance, mex, RNA, elegans, minutes, that, protein, membrane, host, hypothesis, cells, limb, these, may, mitonchondria, aquaponic, been, cenp, ATP, ants, virus, pneumophila, they, F, SHH, chloroplasts, B, treatments, enzymes, cancer, expression, cytochrome, proteins, tularensis, cell, presence, electron, theory, DDT, eukaryotic, O, found*
Key in both	*#, were.*

Economics

Key in Chi-Economics	*rate, model, output, formula, level, growth, curve, income, government, supply, students, population, dividends, per, reserves, consumption, T, aggregate, tax, Dutch, quantity, stock, portfolio, assets, inefficiency, competitive, capm, generation, repurchases, qtmark, asset, refer, cash, disposable, progress, deficit, income*
Key in Eng-Economics	*in, that, as market, however, if, policy, economy, firm, therefore, firms, had, competition, costs, hence, under, since, U.S., significant, period, shown, war, international, lower, world, did, Britain, markets, impact, profits, transport, Bertrand, railways, crises, states, you, cournot, wages, question, extent, stabilisation, British, vertical, shirking, credibility, IMF, governments*
Key in both	*price, demand, monopoly, we, than, capital, increase, higher, exchange, inflation, labour, economic, unemployment, prices, countries, money, production, cost, investment, interest, firm, foreign, crisis, trade, wage, long, marginal, F, country, Y, run, elasticity, domestic, variables, goods, exam, equilibrium, expectations, rates, short, consumers, monetary, surplus, policies, consumer, efficiency, spending, scale, fiscal, productivity, Phillips, slope, bank, central, monopolist, saving, relative*

Engineering

Key in Chi-Engineering	*hydrogen, economy, domain, LC, PVA, directors, cell, BT, tilt, response, director, stylus*
Key in Eng-Engineering	*the, a, is, be, this, for, that, can, was, at, will, have, would, used, results, using, design, been, where, stress, data, beam, then, system, values, analysis, due, value, equation, load, graph, two, pressure, section, force, shows, steel, performance, figure, strain, calculated, shown, ratio, power, factor, point, applied, method, pump, linear, low, current, end, set, car, following, mass, table, voltage, small, appendix, output, line, surface, torque, energy, below, speed, gauges, […] I.*
key in both	*stylus, disc, cantilever, #, amplifier, gauges, drag, moment, mechanical, shaft, gauge, deflection, modelling, discharge, motor, loading, cylinder, capacitor, measurement, bottom, sensor, angle, hole, bridge, analogue, experiment, carbon, circuit, temperature, measured, measuring, efficiency, pressure, length, values, display, digital, resistance, input, head, centre, heat, signal, air, using.*